Abstract

Dependency derivation is the search for combinations of variables (or states of variables) in a database, that co-occur unexpectedly often. In Bayesian dependency derivation, indications are ranked primarily by their estimated strengths, but an adjustment is made to account for uncertainty when data is scarce. This reduces the risk of highlighting spurious associations.

This report presents refined methods for IC analysis—one method for Bayesian dependency derivation. The disproportionality measure in IC analysis is the Information Component (IC) [BLE$^+$98]. It relates the observed joint frequency of two particular states of two different variables to the frequency expected under the assumption of independence.

In the current implementation of IC analysis, estimates for the lower 95% credibility interval limit are derived based on a normal approximation to the posterior IC distribution [OLBL00]. In this report, the validity of these approximations is examined through Monte Carlo simulation. Monte Carlo simulation is also proposed and used as a general tool to study the IC distribution.

For accurate lower credibility interval limit derivation over the entire domain of possible parameter values, two Monte Carlo based approaches are proposed: brute force simulation and a tabular method. These methods vary in execution time and the ranges in which they give accurate results. The optimal combination and implementation of the known approaches is highly dependent on characteristics of the database of interest.

Furthermore, this report shows that for a certain choice of non-informative priors the multinomial and the Poisson data models yield equivalent posterior IC distributions and that Monte Carlo simulation under these circumstances is equivalent to the Bayesian bootstrap.

Relevant aspects of the multiple comparisons issue and problems related to stratification and confounding variables are also discussed.

A Monte Carlo method for Bayesian dependency derivation

Niklas Norén

January 17, 2003

Contents

List of Figures

Part I

Introduction

Chapter 1

Introduction

This is the report of a Master of Science thesis project in Mathematical Statistics, that will conclude a degree in Engineering Physics, at Chalmers University of Technology.

The M.Sc. thesis project has been carried out as part of a joint effort between the research and development company Neurologic in Stockholm and the WHO Collaborating Centre for International Drug Monitoring in Uppsala—also referred to as the Uppsala Monitoring Centre.

The formal examiner of this thesis is Associate Professor Serik Sagitov at the Mathematical statistics department of Chalmers University of Technology. Roland Orre at Neurologic and Stockholm University has been the formal supervisor.

1.1 Background: ADR signal generation in the WHO database

A database held by the Uppsala Monitoring Centre (UMC), in Uppsala, contains more than 2.8 million spontaneous case reports of suspected adverse drug reactions (ADR) [BLE+98]. The database is updated on a quarterly basis with new reports from the 68 member countries of the World Health Organization Programme for International Drug Monitoring. Each case report has 49 fields for e.g. age, sex, ADR, suspected drug substance, concomitant medication, etc. Few reports however carry all this information. All in all there are over 13 000 different drug substances and close to 1900 ADR terms in the database, so the total number of possible drug/ADR combinations is in the order of 10^7.

The purpose of the WHO database is to enable early signaling of drug related adverse events due to drug substances that are already introduced on the market. This is a very important aspect of safety in medicine, since several types of adverse reactions are difficult to identify in clinical trials—e.g rare and long-term side effects, or side effects due to interactions between several drug substances.

To handle today's massive in-flow of data, a method referred to as IC analysis has been implemented on the data set, and has been in routine use since 1998 [BLE$^+$98]. The current implementation is based on a normal approximation to the IC distribution. The aim is to automatically identify the dependencies in the database that are most interesting for a closer investigation. IC analysis is a Bayesian approach, where dependencies are ranked by their lower 95% credibility interval limits for the ratio between the observed joint frequency of two states, and the corresponding expected joint frequency under an assumption of independence.

The rationale for looking at dependencies *within* the database rather than deviations from what can be expected based on sales or prescription data is a lack of reliable such data. In addition, reporting rates may vary for different drug substances and ADR's, in which case any comparison to *external* data will be biased.

In general, it is important to remember that the case reports in the database refer to *suspected* adverse drug reactions that may be due to other circumstances such as e.g. concomitant medication, chance events or the condition for which medication was taken in the first place [Raw88]. Consequently, the naïve ranking of drug/ADR dependencies by shear numbers of reports is biased toward highlighting combinations of the most common drug substances and ADR terms.

1.2 Aims with the thesis

This thesis has three main aims:

- To derive a method for studying the true shape of the IC distribution

- To determine the accuracy of the current normal approximation

- To propose new methods for refined credibility interval estimation, if the current normal approximation can be proven to be inaccurate, for at least some sets of parameters.

1.3 Outline of the thesis

This thesis is divided into four main parts:

Part I These introductory sections.

Part II Recapitulation of and reference to some of the relevant general theory. Focuses especially on Bayesian statistics, Monte Carlo simulation and pseudo-random number generation. Can be browsed through rapidly, or skipped entirely by readers who are already familiar with these areas.

Part III Review of and reference to previous research in IC analysis, as well as in related dependency derivation techniques. A summary of the relevant methods and models that are known today.

Part IV Describes and discusses the implementation of, and the results based, on the new methods introduced in this thesis. In addition, it presents some new conclusions about the method and models in general.

1.4   Acknowledgements

I would like to express my gratitude to the WHO Collaborating Centre for International Drug Monitoring in Uppsala—also known as the Uppsala Monitoring Centre—who funded this thesis project.

I would also like to thank Roland Orre who has been my formal supervisor at Neurologic and Stockholm University, Andrew Bate who has been my contact at the Uppsala Monitoring Centre and Serik Sagitov at the Mathematical Statistics department of Chalmers University of Technology who has been my formal examiner and who also taught the very relevant course in Statistical inference, in the spring of 2002. They have provided excellent guidance and given very valuable feedback to my work.

Finally, I would like to thank friends, family and colleagues, for their support and encouragement.

Stockholm, December 2002,
Niklas Norén.

Part II

Fundamentals

Chapter 2

Review of the relevant theory

As will be explained in Chapter 3, *IC* analysis is a Bayesian method, that greatly emphasizes the posterior *distribution* of the statistic of interest. The new methods proposed in this thesis, are based on Monte Carlo simulation, which in turn relies on pseudo-random number generation. For those who are not familiar with these areas, this chapter gives a brief general introduction. It may be skipped in parts or in its entirety, without loss of context.

2.1 Bayesian statistics

The Bayesian approach is an integral part of *IC* analysis. This section gives a brief overview of the most relevant aspects of Bayesian statistics. For a more thorough treatment, see for example [Lee97].

2.1.1 Bayes theorem

Bayesian statistics combines observed data x and a prior probability distribution $g(\theta \mid prior)$ for the parameter θ, to derive the posterior probability distribution $h(\theta \mid x, prior)$ for θ. The derivation is based on Bayes theorem:

$$h(\theta \mid x, prior) = \frac{f(x \mid \theta, prior)g(\theta \mid prior)}{f(x \mid prior)} \propto f(x \mid \theta)g(\theta \mid prior) \qquad (2.1)$$

The real advantage of Bayesian statistics over the classical approach, is that it allows direct inspection of the *parameter*'s probability distribution. Classical statistics, on the other hand, generally studies the probability distribution of data, and can only draw indirect conclusions about the parameter distribution via the *parameter estimate distribution* (c.f. confidence intervals).

The use of a prior probability distribution is both a strength and a weakness of the Bayesian approach. On one hand, it allows prior knowledge to be incorporated into the analysis, something that may be particularly useful when data is sparse and objective prior information is available. On the other hand, this adds a degree of subjectivity to the analysis, especially when no reliable prior information is available.

This degree of subjectivity is often criticized by statisticians that are skeptical to the Bayesian approach, but it is important to remember that although they are often combined, the Bayesian and the subjectivist views are distinct [HJN89].

2.1.2 Conjugate priors

For a given class F of likelihood functions $f(x \mid \theta)$, the class G of prior distributions $g(\theta)$ is labelled *conjugate* to F if the posterior distribution $h(\theta \mid x)$ is also of class G. With conjugate priors, calculations and derivations of posterior distributions are significantly simplified. It is however crucial to remember that the most important characteristic of a prior distribution is that it adequately describes prior knowledge (or at least prior beliefs).

These are some of the most important likelihood functions and their corresponding conjugate priors:

$\mathbf{f(x \mid \theta)}$	$\mathbf{g(\theta)}$	$\mathbf{h(\theta \mid x)}$
$N(\mu, \sigma^2)$	$N(\mu_0, \sigma_0^2)$	$N\left(\frac{\mu_0 \sigma_0^{-2} + x \sigma^{-2}}{\sigma_0^{-2} + \sigma^{-2}}, \frac{1}{\sigma_0^{-2} + \sigma^{-2}}\right)$
$Bin(n, p)$	$Be(\alpha_0, \beta_0)$	$Be(\alpha_0 + x, \beta_0 + n - x)$
$Mn(n, p_1, \ldots, p_k)$	$Dir(\alpha_1, \ldots, \alpha_k)$	$Dir(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
$Po(\lambda)$	$Ga(\alpha_0, \lambda_0)$	$Ga(\alpha_0 + x, \lambda_0 + 1)$
$Exp(\lambda)$	$Ga(\alpha_0, \lambda_0)$	$Ga(\alpha_0 + 1, \lambda_0 + x)$

2.1.3 Bayesian parameter estimates

The two most common Bayesian parameter estimates are the posterior mean estimate (p.m.e.), which is the mean of the posterior distribution, and the maximum à posteriori (m.a.p.) estimate, which is the value of θ for which the posterior distribution reaches its maximum.

$$\hat{\theta}_{pme} = E(\Theta \mid x) = \int h(\theta \mid x) d\theta \qquad (2.2)$$

$$\hat{\theta}_{map} = \max_{\theta}(h(\theta \mid x)) \qquad (2.3)$$

The posterior mean estimate is sometimes referred to as the Bayes estimate. It minimizes the posterior risk (the expected cost of an erroneous estimate) for squared error loss functions. The maximum à posteriori estimate minimizes the posterior risk for 0-1 loss functions.

2.1.4 Credibility intervals

Credibility intervals are the Bayesian correlate to the confidence intervals of classical statistics. Whereas confidence intervals indicate a most probable interval for data (and indirectly for parameter estimates) under the assumption of a known parameter value, credibility intervals indicate the high probability density region of the parameter θ, given the observed data and prior knowledge [Ric95].

2.2 Monte Carlo simulation methods

This section, gives a brief review of Monte Carlo simulation methods, and in particular of different versions of the bootstrap method.

2.2.1 Monte Carlo simulation

Even the values of seemingly simple functions of random variables may follow probability distributions that are difficult to analyze. Over-simplified model assumptions are therefore often made in order to make possible the evaluation of non-standard distributions or statistics. For example, dubious normal approximations may be erroneously accepted because there is no known analytical way to calculate confidence/credibility intervals for a particular estimate/parameter. In other situations, a known model of the dependency between trials may be ignored so that results for independent identically distributed (i.i.d.) trials will apply.

Monte Carlo simulation is an alternative to deliberate errors. The basic strategy is simple: draw a large number of random samples from the distribution/model of interest and use the simulated distribution to make inference about the real distribution (parameter estimates, confidence/credibility intervals etc.). A smoothed version of the simulated cumulative distribution function will tend to the true cumulative distribution function as the number of draws tends to infinity. Thus, any conclusions about the true distribution based on the simulated distribution will get arbitrarily accurate as the number of draws increase.

2.2.2 Bootstrap methods

The bootstrap is a generalization of Monte Carlo simulation, that is used when the true distribution to be simulated is unknown. With a bootstrap method, the simulation is based on a distribution estimate inferred from data.

2.2.3 The non-parametric bootstrap

In the non-parametric bootstrap, new batches of data, the same size as the original batch, are simulated by sampling with replacement from the original

data batch. This corresponds to sampling with replacement from the empirical cumulative distribution function F_n [Ric95].

The statistic of interest is calculated for each simulated batch of data, and the simulated distribution of the statistic is studied to draw conclusions about the distribution of the true statistic. The advantage of the non-parametric bootstrap over the parametric bootstrap described in Section 2.2.4 is that the non-parametric bootstrap makes no assumption about the underlying model.

Its accuracy is on the other hand difficult to evaluate since it depends both on how well F_n approximates F (the true cumulative distribution) and on how sensitive the investigated statistic is to variations in F [Ric95].

2.2.4 The parametric bootstrap

When particular model assumptions are motivated, a parametric version of the bootstrap is often used. In the parametric bootstrap, an approximate cumulative distribution is found by estimating from data the parameters of an assumed model. New data batches (of the same size as the original batch) are then generated from the parameterized distribution estimate, and the statistic of interest is calculated for each simulated batch of data.

The accuracy of the parametric bootstrap is generally rather sensitive to a correct model choice. In addition, its accuracy depends on the accuracy of the parameterized distribution estimate.

2.2.5 The Bayesian bootstrap

In the Bayesian bootstrap [Rub81] the original data set is resampled by assigning a random weight to each observation in such a way that the weights follow the n-dimensional Dirichlet distribution $Dir_n(1, 1, 1, 1, ...)$ where n is the size of the original data set.

This assignment of random weights corresponds to the resampling of the regular bootstrap which can be equivalently described as drawing random weights from the discrete set $\{0, \frac{1}{n}, \ldots, \frac{n}{n}\}$. The main difference between the regular and the Bayesian bootstrap is in fact in how the random weights are assigned [CL01]. (Although, there is of course also a difference in interpretation, since the Bayesian bootstrap simulates the posterior *parameter distribution* and the regular bootstrap simulates the *parameter estimate distribution*)

One advantage of the Bayesian bootstrap over the standard bootstrap is that even though individual samples may be assigned very low weights, they are never completely excluded from the resampled data set. This eliminates the risk of parameter estimates that are inconsistent with the observed data (for example $\hat{p} = 0$ given $X = x > 0$ in a Bernoulli experiment), typical for the regular bootstrap.

2.3 Pseudo-random number generation

The area of pseudo-random number generation has been subject to extensive research in recent years, mainly due to the increased use of simulation methods in e.g. mathematics, physics and biology. Random numbers are of fundamental importance to Monte Carlo simulation since without proper randomization, correct simulation is impossible.

2.3.1 Uniformly distributed pseudo-random numbers

Most computers include software for uniform random number generation, but these numbers are not truly random. Rather, the standard pieces of software generate sequences of *pseudo-random* numbers that imitate true i.i.d. random numbers very well. These sequences often pass all statistical tests for random sequences, but they *are* nevertheless generated through deterministic procedures [Häg02].

Whether truly random or not though, the pseudo-random numbers of today serve their purpose well, and uniform random number generation is the foundation on which more general random number generation relies.

2.3.2 The inversion method

The inversion method is a straightforward strategy for simulation of non-uniform random variates. It is applicable to any function that has a strictly increasing (to avoid ambiguities) cumulative distribution function (c.d.f.) F with a known inverse F^{-1}.

The underlying theory is fairly simple: a random cumulative distribution value $F(X = x)$ is simulated by drawing a uniform random variable Y on $[0, 1]$; due to the 1:1 correspondence between $F(X)$ and X (since $F(X)$ is strictly increasing) the uniform random variable y can be transformed to the more general random variable x through inversion, $x = F^{-1}(y)$.

2.3.3 The rejection method

Often, no closed form expression for the inverse of the c.d.f. is available. The rejection method enables random number generation from any probability density function $p(x)$, provided that there is a constant A and an envelope function $q(x)$ (whose random variates should be easy to generate) so that $p(x) < Aq(x)$ for all x. (If $p(x)$ is upper bounded by a constant, the uniform distribution is a possible envelope function.)

The rejection method consists of drawing one random number $\xi \sim q$ and another random number $u \sim U[0, 1]$. ξ is accepted if $Auq(\xi) \leq p(\xi)$, and it can be shown that $\xi_{acc} \sim p(x)$ [Sch81].

Part III

Overview of known methods and models

Chapter 3

The IC analysis methodology

The strategy in IC analysis is to rank dependencies in a database primarily with respect to their indicated strengths but with an adjustment to account for uncertainty, in order to reduce the risk of highlighting spurious associations.

3.1 The Information Component

The Information Component (IC) is defined between any two states of variables in a database (or a similar data set). If the states of interest are $X = x$ and $Y = y$, the IC can be expressed as [KO98]:

$$IC = \log_2 \frac{P(Y = y \mid X = x)}{P(Y = y)} \tag{3.1}$$

Equivalent expressions are:

$$IC = \log_2 \frac{P(X = x \mid Y = y)}{P(X = x)} \tag{3.2}$$

and:

$$IC = \log_2 \frac{P(X = x, Y = y)}{P(X = x)P(Y = y)} \tag{3.3}$$

With respect to the two states of interest, there are at most four different types of records in a database: records with none of the two states, records with either one of the two states and records with both states. The total numbers of records of each type are in this report denoted: c_{00}, c_{01}, c_{10} and c_{11} respectively. The corresponding frequencies are denoted f_{00}, f_{01}, f_{10} and f_{11}, and the assumed underlying probabilities are denoted p_{00}, p_{01}, p_{10} and p_{11}. The marginal probabilities of the two states are denoted $p_{1.}$ and $p_{.1}$ etc. The annotation for the marginal counts and frequencies is equivalent.

With this annotation, the IC is:

$$IC = \log_2 \frac{p_{11}}{p_{1 \cdot} p_{\cdot 1}} \tag{3.4}$$

The IC was originally derived as the weight between two nodes in a Bayesian neural network [BLE+98] and it is closely related to Mutual Information (see Section 4.3). The sign of the IC indicates whether the joint probability of the two states is $(+)$ greater or $(-)$ smaller than what is expected under the assumption of independence between the two states, and the absolute value of the IC indicates how strong the dependency between the two states are.

To illustrate that such a statistic may be of interest in applications other than drug monitoring, consider the following example from market basket analysis: the observation that milk and flour are often purchased together may be of limited interest if this is fully explained by the fact that milk is overall very frequently purchased. Because of the term $p_{1 \cdot} p_{\cdot 1}$ in the denominator, the IC only highlights dependencies that are not attributable to high marginal support alone.

A naïve IC point estimate from classical statistics is:

$$IC = \log_2 \frac{f_{11}}{f_{1 \cdot} f_{\cdot 1}} \tag{3.5}$$

but this may be biased and does not account for uncertainty.

3.2   The Bayesian approach

Bayesian statistics is an important part of the IC analysis methodology. With a Bayesian approach, analysis of the posterior IC distribution allows uncertainty to be accounted for. In IC analysis, dependencies are ranked by their estimated lower credibility interval limits (denoted $IC_{\alpha/2}$) for a given significance level α.

An alternative approach would be to rank dependencies by the proportion of the posterior IC distribution that exceeds a certain threshold β: $P(IC > \beta)$. The $IC_{\alpha/2}$ ranking, however, has the advantage of ranking dependencies primarily by their indicated relative strengths. As the amount of information on a dependency increases, $IC_{\alpha/2}$ will tend to the point estimates of the IC.

In the current implementation, dependencies are ranked by $IC_{0.025}$ ($\alpha = 5\%$).

3.3   Known data models

The database is considered to be a random sample from an underlying population that is the true focus of the study. At least three different models can be used to describe the distribution of records in the database.

For a review of the relevant probability distributions, please see Appendix A.

3.3.1 The bB data model

The model on which the current implementation of IC analysis is based, considers the distribution of data to be determined by three binomial distributions. $c_{11} \sim bin(n, p_{11})$, $c_{1\cdot} \sim bin(n, p_{1\cdot})$ and $c_{\cdot 1} \sim bin(n, p_{\cdot 1})$ [OLBL00]. The use of three independent distributions does not allow covariation between the parameters to be accounted for.

Since the Beta distribution is the conjugate prior (see Section 2.1.2) of the binomial distribution the prior distributions for p_{11}, $p_{1\cdot}$ and $p_{\cdot 1}$ in this model are usually selected from the Beta distribution family. The combined model is therefore referred to as the binomial/Beta data model (or the bB data model, for short).

3.3.2 The PG data model

The Gamma-Poisson shrinkage method [DuM99] is an approach similar in spirit to IC analysis. As the name suggests, data is in this method assumed to follow a Poisson distribution. In particular, the database is modelled as the sum of four independent Poisson distributions—one for each type of report, with intensities: λ_{00}, λ_{01}, λ_{10} and λ_{11}. In this model, the size of the database, $c_{\cdot\cdot}$ is not assumed to be fixed in advance, but rather to be the sum of the random counts: c_{00}, c_{01}, c_{10} and c_{11}.

Since the Gamma distribution is the conjugate prior (see Section 2.1.2) of the Poisson distribution, the prior distributions for λ_{00}, λ_{01}, λ_{10} and λ_{11} in this model are usually selected from the Gamma distribution family. The combined model is therefore referred to as the Poisson/Gamma data model (or the PG data model, for short).

3.3.3 The mD data model

The possibility to model the distribution of counts in the database by a joint multinomial probability distribution rather than by separate binomial distributions is mentioned in [OLBL00]. It has however not been fully pursued in earlier research.

In the multinomial model, the set of counts $\{c_{00}, c_{01}, c_{10}, c_{11}\}$ is assumed to follow a $Mn(p_{00}, p_{01}, p_{10}, p_{11}, c_{\cdot\cdot})$ distribution. This means that the marginal probabilities of c_{11}, $c_{1\cdot}$ and $c_{\cdot 1}$ follow the same distributions as in the bB data model, but an advantage of the multinomial model is that it allows covariation between the marginal counts to be properly accounted for.

Since the Dirichlet distribution is the conjugate prior (see Section 2.1.2) of the multinomial distribution, the prior distribution for $\{p_{00}, p_{01}, p_{10}, p_{11}\}$ in this model is usually selected from the family of Dirichlet distributions. Again, with respect to the marginal probabilities, this corresponds perfectly to the bB data model, since the marginal probabilities of a joint Dirichlet distribution, are Beta distributed. This model is referred to as the multinomial/Dirichlet data model (or the mD data model, for short).

3.4 Known prior distributions

IC analysis is a Bayesian method, and an important design parameter is consequently the choice of prior distributions for different parameters. This section assumes either the bB or the mD data models, and for clarity, the following general annotation is used:

$\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11}$ Refer specifically to the prior distribution parameters
$\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}$ Refer specifically to the posterior distribution parameters

Here, $\alpha_{1.} = \alpha_{11} + \alpha_{10}$ etc. as before.

It is worth noting that in the bB and the mD data models, the posterior parameters are equal to the sum of the counts and the prior parameters ($\gamma_{ij} = c_{ij} + \alpha_{ij}$). The prior parameters may therefore be referred to as pseudo-counts.

In addition, the sum of the prior parameters $\alpha_{..} = \alpha_{00} + \alpha_{01} + \alpha_{10} + \alpha_{11}$ is generally referred to as the *equivalent sample size* of the prior. It indicates how important prior information is considered to be relative to real data.

3.4.1 The current prior distribution

These are the prior distributions used in the current implementation of the method:

$$p_{1.} \sim Be(1, 1)$$
$$p_{.1} \sim Be(1, 1)$$
$$p_{11} \sim Be(1, \frac{c_{..}^2}{c_{1.}c_{.1}} - 1)$$

The main motivation for this choice of prior distributions is that they indicate independence for any unobserved combination, by ascertaining that [BLE$^+$98]:

$$\lim_{c_{11} \to 0} IC = 0 \tag{3.6}$$

This choice may however be criticized on the basis of not being truly *prior*, since the prior distribution for p_{11} depends directly on observations $c_{1.}$ and $c_{.1}$ in the database.

In addition, the equivalent sample sizes of this prior have the odd characteristic of being different for the different parameters (2 for $p_{1.}$ and $p_{.1}$, and $\frac{c_{..}^2}{c_{1.}c_{.1}}$ for c_{11}) as well as for different combinations (inversely proportional to the product of $c_{1.}$ and $c_{.1}$). This means that the impact of prior information does not only vary between combinations, but also between p_{11} and $p_{1.}$ / $p_{.1}$. The equivalent sample size of the prior for p_{11} is generally much larger than 2, but there is no obvious reason why more impact should be attributed to prior information for p_{11} than for $p_{1.}$ or $p_{.1}$.

3.4.2 Other informative prior distributions

There is of course an infinite number of possible informative prior distributions, since for example in the mD data model, all four parameters of a joint Dirichlet prior distribution may take on any non-negative value. Different approaches to the derivation of informative priors may however be discussed.

One strategy for derivation of sensible prior distributions is to use information from outside of the database (e.g. sales data or demographical data).

Another possibility is to carry out empirical Bayes estimation. In empirical Bayes estimation, an informative prior distribution is derived by fitting a prior distribution model to the observed data (previous or current) [Mar70]. In our application, a Beta or a Dirichlet distribution could be fitted to the frequencies for a large number of combinations in the database that are assumed to follow the same prior distribution. This method has been used for a statistic similar to the IC [DP01]. Because prior distributions derived through empirical Bayes estimation depend on data, empirical Bayes estimation is sometimes referred to as a semi-Bayesian method.

3.4.3 Non-informative prior distributions

A common way to handle lack of prior knowledge is to use a so-called non-informative prior. The most important aspect of a non-informative prior is that it has minimal influence on the posterior distribution. For Beta distributions, at least three non-informative priors have been proposed: Haldane's prior $Be(0,0)$, the arc-sine prior $Be(1/2, 1/2)$ and the uniform prior $Be(1,1)$ [Lee97].

Haldane's prior is improper (it does not integrate to 1) but it has the advantage of yielding the most data sensitive posterior distribution [Lee97], as well as generalizing properly to Dirichlet and Gamma prior distributions ($Dir(0,0,\ldots,0)$ and $Ga(0,0)$ may be used). In addition, the posterior distributions will always be proper, except for combinations with non-zero counts c_{11} in the database.

The main drawback of non-informative priors is that the posterior IC distributions may vary largely for small changes in low counter values. Analysis based on non-informative priors is therefore prone to highlighting spurious associations when data is sparse. A certain state of a variable could in fact be overall so rare in a large database, that based on the marginal frequencies, it is unlikely to occur with *any* one state of another variable!

3.5 Known approximations to the IC distribution

Although, the explicit posterior distributions for the constituting parameters p_{11}, $p_{1\cdot}$ and $p_{\cdot 1}$ are known, there are currently no methods for exact analytical evaluation of the posterior IC distribution [OLBL00]. Consequently IC analysis relies on approximate methods.

3.5.1 The currently implemented normal approximation

In the current implementation, the derivation of credibility interval limits is based on a normal approximation to the IC distribution. In addition to the general approximation in replacing the true distribution by a normal distribution, the formulae for the mean and variance are approximate as well.

The current approximation of the p.m.e. is [OLBL00]:

$$E(IC) = E(\log_2 p_{11}) - E(\log_2 p_{1\cdot}) - E(\log_2 p_{\cdot 1}) \approx \qquad (3.7)$$
$$\approx \log_2 E(p_{11}) - \log_2 E(p_{1\cdot}) - \log_2 E(p_{\cdot 1}) =$$
$$= \log_2\left(\frac{\gamma_{11}\gamma_{\cdot\cdot}}{\gamma_{1\cdot}\gamma_{\cdot 1}}\right)$$

The current approximation of the variance is [OLBL00] (to simplify annotation, $\hat{p}$ here denotes the $\hat{p}_{map}$, which for a $Be(\alpha,\beta)$ distribution is: $\frac{\alpha}{\alpha+\beta}$):

$$Var(IC) \approx \frac{Var(p_{11})(\frac{1}{\hat{p}_{11}})^2 + Var(p_{1\cdot})(\frac{-1}{\hat{p}_{1\cdot}})^2 + Var(p_{\cdot 1})(\frac{-1}{\hat{p}_{\cdot 1}})^2}{(\ln 2)^2} = \qquad (3.8)$$
$$= \frac{\frac{\hat{p}_{11}(1-\hat{p}_{11})}{(c_{\cdot\cdot}+1)(\hat{p}_{11})^2} + \frac{\hat{p}_{1\cdot}(1-\hat{p}_{1\cdot})}{(c_{\cdot\cdot}+1)(\hat{p}_{1\cdot})^2} + \frac{\hat{p}_{\cdot 1}(1-\hat{p}_{\cdot 1})}{(c_{\cdot\cdot}+1)(\hat{p}_{\cdot 1})^2}}{(\ln 2)^2} =$$
$$= \frac{\frac{(1-\hat{p}_{11})}{\hat{p}_{11}} + \frac{(1-\hat{p}_{1\cdot})}{\hat{p}_{1\cdot}} + \frac{(1-\hat{p}_{\cdot 1})}{\hat{p}_{\cdot 1}}}{(c_{\cdot\cdot}+1)(\ln 2)^2}$$

Equation 3.8 is based on a Gauss approximation that disregards covariation and combines:

$$z = y_1 + y_2 + \ldots + y_k \qquad \Rightarrow \qquad s_z^2 = s_{y_1}^2 + s_{y_2}^2 + \ldots + s_{y_k}^2 \qquad (3.9)$$
$$y = log_2(x) \qquad \Rightarrow \qquad s_y^2 = \left(\frac{s_x}{x\ln 2}\right)^2 \qquad (3.10)$$

with the expression for the variance of $p \sim Be(\alpha,\beta)$:

$$Var(p) = \frac{\hat{p}(1-\hat{p})}{\alpha+\beta+1} \qquad (3.11)$$

3.5.2 A refined normal approximation

More precise estimates of the mean and the variance give a more accurate normal approximation. Based on the following result for $p \sim Be(\alpha,\beta)$ derived in [OLBL00] the posterior mean can be calculated exactly:

$$E(\log p) = \frac{\beta}{\alpha(\alpha+\beta)} - \beta\sum_{i=1}^{\infty}\frac{1}{(\alpha+i)(\alpha+\beta+i)} \qquad (3.12)$$

The exact expression for the variance of the IC distribution is as follows [OLBL00]:

21

$$Var(IC) = Var(\log p_{11}) + Var(\log p_{1\cdot}) + Var(\log p_{\cdot 1}) + \qquad (3.13)$$
$$- 2Cov(\log p_{11}, \log p_{1\cdot}) - 2Cov(\log p_{11}, \log p_{\cdot 1}) +$$
$$+ 2Cov(\log p_{1\cdot}, \log p_{\cdot 1})$$

The terms $Var(\log p_{11})$ etc. can be evaluated exactly based on the following expression for $p \sim Be(\alpha, \beta)$ [OLBL00]:

$$Var(\log p) = \sum_{i=0}^{\infty} \frac{\beta^2 + 2\alpha\beta + 2\beta i}{(\alpha + i)^2 (\alpha + \beta + i)^2} \qquad (3.14)$$

However, no similar formulae for the covariances are known, so there is currently no exact formula for $Var(IC)$.

3.5.3 A fixed marginals approximation

A different approximation is implicit in the Gamma-Poisson shrinkage method and its successor the multi-item Gamma-Poisson shrinkage method [DuM99, DP01]. These methods are based on a statistic similar to the IC that only accounts for uncertainty in p_{11}. A similar approximation to the IC distribution would be:

$$IC_{fix} = \log_2 \frac{p_{11}}{\hat{p}_{1\cdot_{map}} \hat{p}_{\cdot 1_{map}}} \qquad (3.15)$$

In this approximation, the IC is simply the logarithm of a Beta distributed random variable (p_{11}) divided by a constant $(\hat{p}_{1\cdot}\hat{p}_{\cdot 1})$. Since the incomplete Beta function can be readily evaluated, a Newton-Raphson iterative method (see for example [Hea97]) may be used to find the value of p_{11} for which the incomplete Beta function is equal to α—this is the lower credibility interval limit for p_{11} $(\hat{p}_{11_{\alpha/2}})$. And because $log(p_{11})$ is monotone in p_{11}, the $\hat{IC}_\alpha$ estimate based on the fixed marginals approximation is simply:

$$\hat{IC}_\alpha = \log_2 \frac{p_{11_{\alpha/2}}}{\hat{p}_{1\cdot_{map}} \hat{p}_{\cdot 1_{map}}} \qquad (3.16)$$

Chapter 4

Other methods for dependency derivation

Dependency derivation is the search for dependent variables or states of variables in a database or another data set. It is an important matter both in traditional statistical inference analysis, and in an emerging field of research referred to as knowledge discovery in databases (KDD) [FPSS96].

KDD combines methods and results primarily from statistical inference theory, machine learning theory and database theory. Its scope includes the collection, storage, maintenance, extraction and cleaning of data [Vea02]. One out of several proposed definitions is: 'the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data' [FPSS96].

Of the types of methods presented in this chapter, the first is a typical statistical inference technique, and the second a typical KDD technique. The third section describes a statistic relevant to dependency derivation, that has its origin in Shannon's information theory.

4.1 Tests for independence

A common statistical inference technique used to determine whether two variables are independent or not, is to construct contingency tables and perform either a chi-squared test for independence (when all expected counts are larger than four) or a Fisher's exact test (when some expected counts are lower than five).

The drawback with hypothesis testing is that it tends to emphasize precision over relevance, i.e. slight effects with precise estimates are favored over less certain indications of stronger effects. Often, the focus in dependency derivation is not just the statistical significance of the dependency, but also the strength (whether a certain combination occurs 10 or 1.1 times more often than expected under the assumption of independence).

4.2 Association rule analysis

Association rule analysis is a KDD method for dependency derivation.

The general form of an association rule is [HTF01]:

$$\{X_1 = x_1, ..., X_m = x_m\} \Rightarrow \{Y_1 = y_1, ..., Y_n = y_n\} \qquad (4.1)$$

or with simplified notation:

$$A \Rightarrow B \qquad (4.2)$$

Generally speaking, the objective of Association rule analysis is to find different types of tendencies in the database, e.g. association rules with high confidence or high lift (see below!).

The 'support' of an association rule is defined as:

$$P(A) \qquad (4.3)$$

the 'confidence' as:

$$P(B \mid A) \qquad (4.4)$$

and the 'lift' as:

$$\frac{P(B \mid A)}{P(B)} \qquad (4.5)$$

High confidence indicates that the probability of event B is high given that event A is true. High lift on the other hand means that the probability of event B is higher if event A is true.

For computational efficiency, many of the association rule analysis algorithms (e.g. the Apriori algorithm [HTF01]) reduce search space by only considering the combinations of variables that have high marginal support (in a greedy algorithm fashion). This is based on an assumption that to be interesting a dependency must have large support—something that may be true for sales data but not necessarily in other applications (in the WHO database for example, early signaling is very important). The restriction to combinations with large support automatically makes impossible the identification of significant dependencies between rare states of variables in the database.

The main drawback of association rule analysis is however that variability in the estimates is completely unaccounted for: no distinction is made between a lift of 2 based on 10 observations and a lift of 2 based on 1000 observations. This is the opposite of the drawback with hypothesis tests: relevance but not significance is accounted for. Association rule analysis is consequently prone to spurious associations.

4.3 Mutual Information

Mutual Information is a statistic, that originates from Information theory and measures the strength of the dependency between two variables. The Mutual Information between two variables X and Y is defined as [CT91]:

$$I_{XY} = \sum_{x,y} p_{xy} \log \frac{p_{xy}}{p_{x \cdot} p_{\cdot y}} \qquad (4.6)$$

One way to look at this expression for Mutual Information, is as a weighted sum of the Information Components between all possible combinations of states for the two variables.

Mutual Information is a well-known statistic for which there are known and accurate approximations to the posterior mean and the variance [Hut01]. This speaks in favor of a dependency derivation approach based on Mutual Information rather than on the IC, but an important advantage of the IC is its specificity—when it is the possible dependency between two specific *states* of two variables that is of interest, the use of Mutual Information may be misleading.

Consider two binary variables X and Y and assume that we are interested in whether $X = 1$ often co-occurs with $Y = 1$ (this may represent the co-occurrence of a specific drug and ADR combination on a report in the WHO database). Consider the corresponding aggregated data sets (IC_{00} refers to the Information Component between the states $X = 0$ and $Y = 0$ and similarly for IC_{01} and IC_{10}):

	Set I	Set II
γ_{00}	10 912	20 912
γ_{01}	27	27
γ_{10}	57	57
γ_{11}	4	4
IC_{00}	0.00051	0.00027
IC_{01}	-0.19	-0.20
IC_{10}	-0.09	-0.10
IC_{11}	4.54	5.47
I_{XY}	0.0012	0.0008

Note that $I_{XY}^I > I_{XY}^{II}$, even though $IC_{11}^I < IC_{11}^{II}$! The explanation for this is that the Information Components except for the one of interest, IC_{11}, is lower in configuration II. In addition, the combination of interest is very rare in the database and therefore has little impact on the Mutual Information calculation where each Information Component IC_{ij} is weighted by p_{ij}.

Clearly in an application where we are particularly interested in the dependency between two specific states of two variables, the proper Information Component is a more appropriate measure than Mutual Information.

Another advantage of IC in is that its sign indicates whether the co-occurrence of the two states is unexpectedly common or unexpectedly rare. The Mutual Information between two variables is on the other hand always greater than 0, and enables no such distinction.

Part IV

Monte Carlo analysis of the IC distribution

Chapter 5

Methods and model

The main contribution of this thesis to ongoing research in IC analysis is the incorporation of a Monte Carlo method for more accurate information about the IC distribution, and as an alternative to the approximate methods currently used.

Principles of the method and results based on it are presented and discussed in this final and most important part of the report. All simulations are implemented and carried out in the MATLAB environment.

5.1 Model assumptions and annotation

The implementation of the Monte Carlo method is based on a mD (multinomial/Dirichlet) data model (see Section 3.3.3). Specifically, the following assumptions are made about data and parameter distributions:

$$\{c_{00}, c_{01}, c_{10}, c_{11}\} \sim Mn(p_{00}, p_{01}, p_{10}, p_{11}, c_{..})$$
$$\{p_{00}, p_{01}, p_{10}, p_{11}\}_{prior} \sim Dir(\alpha_{00}, \alpha_{01}, \alpha_{10}, \alpha_{11})$$
$$\{p_{00}, p_{01}, p_{10}, p_{11}\}_{posterior} \sim Dir(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$$

Since the analysis of the IC *prior* distribution is similar to that of the IC *posterior* distribution, the results presented in the following sections apply to both. We will however, for simplicity, refer to the parameters of a general IC distribution as γ_{ij}.

As mentioned earlier, the γ parameters of the mD model are closely related to the database counts. For a prior distribution, they may be thought of as pseudo-counts representing prior knowledge, and for a posterior distribution, they are the sum of the true counts and the pseudo-counts. If Haldane's prior (see Section 3.4.3) is used, the pseudo-counts are all 0, and the γ parameters are consequently equal to the database counts. The Monte Carlo method as such, is however independent of the choice of prior distribution.

5.2 Setup of the systematic analysis of the IC distribution

The values of the four parameters γ_{00}, γ_{01}, γ_{10} and γ_{11} in the joint Dirichlet distribution of p_{00}, p_{01}, p_{10} and p_{11} all have an impact on the IC distribution. Since they may take on any non-negative values, exhaustive, simulation-based evaluation of the shape of the IC distribution for all possible sets of parameter values is intractable. The results in Chapter 6 are therefore based on separate studies of each parameter's individual impact. The validity of this setup is commented on in the discussion (see Section 7.2).

Since both prior and posterior distribution parameters in the mD model can be thought of as pseudo-counts (see Section 3.4), and since the database itself is often described in terms of c_{11}, $c_{1\cdot}$, $c_{\cdot 1}$ and $c_{\cdot\cdot}$, we will generally consider the corresponding parameters γ_{11}, $\gamma_{1\cdot}$, $\gamma_{\cdot 1}$ and $\gamma_{\cdot\cdot}$ rather than the fundamental parameters γ_{00}, γ_{01}, γ_{10} and γ_{11} in the study of parameter impact on the IC distribution. Of course when all parameters but one a are kept fixed, there is a $1:1$ correspondence between these aggregated parameters and the fundamental parameters: exclusive variation in $\gamma_{1\cdot}$ corresponds to variation in γ_{10}, exclusive variation in $\gamma_{\cdot 1}$ corresponds to variation in γ_{01} and exclusive variation in $\gamma_{\cdot\cdot}$ corresponds to variation in γ_{00}.

To enable quantitative examination of the shape of the distribution, the following features are measured: standard deviation (the spread), skewness (the degree of asymmetry) and kurtosis (tendency to be heavy-tailed). Standard deviation and kurtosis range from 0 to infinity and skewness may take on any real value. For a normal distribution the skewness is 0 and the kurtosis is 3. A negative skewness indicates a distribution skewed to the left, and a positive skewness indicates a distribution skewed to the right. A kurtosis larger than 3 indicates tails heavier than those of the normal distribution and a kurtosis smaller than 3 indicates tails that are lighter.

5.3 Implementation of the Monte Carlo method

The Monte Carlo method for analysis of the IC distribution has been successfully implemented in a number of MATLAB routines. If used in routine IC analysis of the WHO database, the MATLAB routines will be replaced by more efficient programs in a compiled language such as C.

5.3.1 Outline of the method

The general strategy of the Monte Carlo method for IC analysis is rather straightforward:

1. Draw a large number of parameter sets $\{p_{00}^*, p_{01}^*, p_{10}^*, p_{11}^*\} \sim Dir(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$

2. For each randomly drawn parameter set, calculate $IC^* = \log_2 \frac{p_{11}^*}{p_{1\cdot}^* . p_{\cdot 1}^*}$
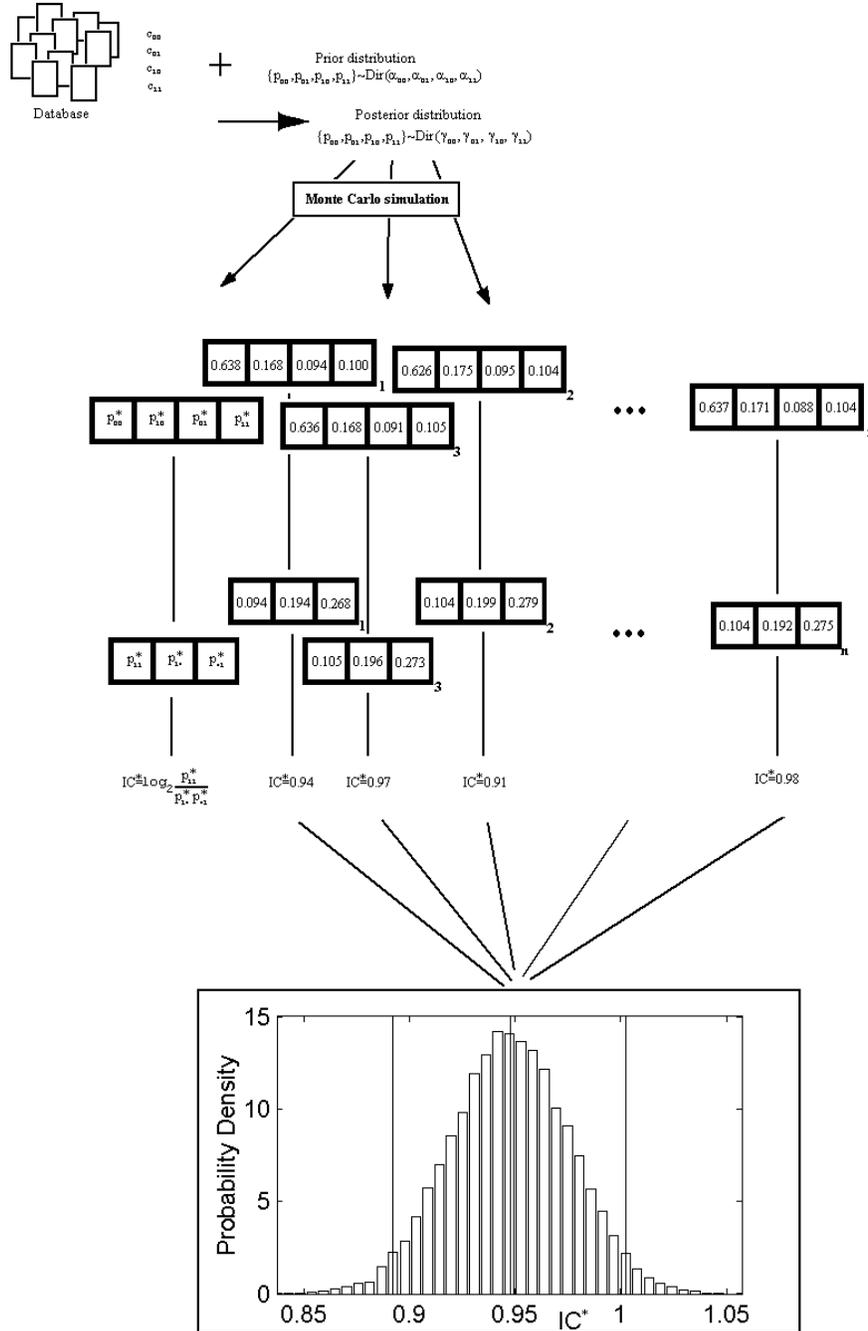
Figure 5.1: Schematic description of the Monte Carlo method for IC analysis

3. Study the IC^* distribution to make inference about the IC distribution

Credibility intervals in general, and $IC_{\alpha/2}$ (see Section 3.2) in particular, are estimated by the corresponding simulation quantiles. (As an aside, note that the question of how to best estimate confidence intervals based on classical bootstrap simulation is much more heavily debated, and that no method proposed so far has been generally accepted to be superior to the others [Hjo94].)

Ideally, a closed form expression for the accuracy of the Monte Carlo-based $IC_{\alpha/2}$ estimation, would indicate the number of draws necessary for a certain precision in this estimate, but no such expression is known.

Instead, the variation in $\hat{IC}_{\alpha/2}$ for a given number of draws, can be estimated by studying its empirical variance over a large number of repeated simulations. This result will be specific to that particular combination of distribution shape and number of draws in the simulation. Since the uncertainty in Monte Carlo-based quantile estimates is proportional to the spread of the simulated distribution, an upper limit for the variation can be found by considering the sampling variability for estimates based on the most spread-out distribution.

The current normal approximation indicates that the spread of the IC distribution increases with decreased values for all the parameters of the Dirichlet distribution. An initial experiment with 100 simulations (each with 30 000 draws) of $IC_{\alpha/2}^*$, was carried out for a $Dir(1,1,1,1)$ parameter distribution. In this experiment, all 100 $IC_{\alpha/2}^*$ estimates were within an interval more narrow than 0.2 bits. It is consequently a reasonable assumption that the accuracy of $IC_{\alpha/2}^*$ based on 30 000 draws is at the very least better than 0.2 bits (and generally much better).

The benefit of the Monte Carlo method is that for a large enough number of draws, the IC^* distribution is a precise and unbiased approximation to the IC distribution. The drawback with this method is that it is computationally intensive, and that the computational complexity increases with the number of draws (thus with the demand for accuracy).

5.3.2 Dirichlet random variate generation

In Monte Carlo simulation of the IC distribution, random configurations: $\{p_{00}^*, p_{01}^*, p_{10}^*, p_{11}^*\} \sim Dir(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$ need to be generated.

This is accomplished by simulating a carefully selected set of marginal probabilities: $P(X = x)$, $P(Y = y \mid X = x)$ and $P(Y = y \mid X \neq x)$. These marginal probabilities follow Beta distributions for which random variates can be generated with Cheng's BB algorithm [Che78], which is a type of rejection method (see Section 2.3.3):

$$
\begin{aligned}
p_{1.}^* &\sim Be(\gamma_{1.}, \gamma_{..} - \gamma_{1.}) \\
p_{11|1.}^* &\sim Be(\gamma_{11}, \gamma_{1.} - \gamma_{11}) \\
p_{01|0.}^* &\sim Be(\gamma_{01}, \gamma_{0.} - \gamma_{01})
\end{aligned}
\tag{5.1}
$$
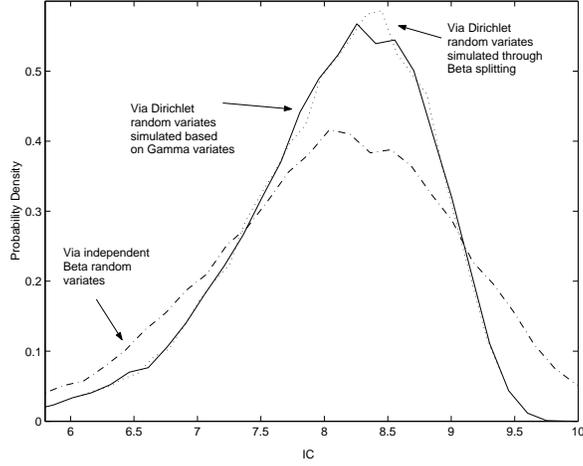
Figure 5.2: Comparison of IC simulations based on the binomial/Beta data model and on the mD data model (either Gamma or Beta based). $\gamma_{11} = 3$, $\gamma_{1.} = 10$, $\gamma_{.1} = 10^6$ and $\gamma_{..} = 10^6$. 30 000 draws were used in each simulation.

The most important property of this set of marginal probabilities is that although they have been simulated independently, these marginals can be used to reconstruct the joint Dirichlet distribution without failing to account for covariation.

The Dirichlet random variates are:

$$
\begin{aligned}
p_{00}^* &= (1 - p_{01|0.}^*)(1 - p_{1.}^*) \\
p_{01}^* &= p_{01|0.}^*(1 - p_{1.}^*) \\
p_{10}^* &= (1 - p_{11|1.}^*)p_{1.}^* \\
p_{11}^* &= p_{11|1.}^* p_{1.}^*
\end{aligned}
$$

$$(5.2)$$

To illustrate the importance of accounting for covariation in IC simulation, and to underline that this is properly accomplished with the proposed Dirichlet random generator, Figure 5.2 compares simulated IC distribution based on this and two other random generators: one for the bB data model (based on independent Beta random variates for p_{11}, $p_{1.}$ and $p_{.1}$) and one for the mD model (based on Dirichlet random generation through Gamma variates).

The similarity in this figure between the Gamma and the Beta based simulations supports the validity of the proposed method. The noticeable difference between these two and the IC distribution based on the bB data model, on the other hand illustrates the impact that neglected covariation may have on the

IC distribution. In fact, since marginal distributions often overlap, bB based simulation may even produce inconsistent triplets where e.g. $p_1. < p_{11}$ (which is obviously incorrect since $p_1. \equiv p_{11} + p_{10} \geq p_{11}$).

The motivation for using the less straightforward Beta type Dirichlet random generator rather than the standard Gamma type, is that for a four dimensional Dirichlet distribution, the Beta method seems to be more efficient.

5.4 Method related issues

This section is a compilation of method and model related conclusions, that to my knowledge have not been mentioned in previous research.

5.4.1 On covariation in the mD data model

The advantage of the full mD data model over the simplified bB data model, is that the former allows for proper derivation of covariation between p_{11}, $p_1.$ and $p._1$. This section is devoted to find expressions for these covariances. These expressions will turn out to be very helpful later on in the analysis of how (and when) the different parameters affect the shape of the *IC* distribution.

As an aside, note that the variances of the parameters are those of the corresponding Beta distributions:

$$Var(p_{11}) = \frac{\hat{p}_{11}(1 - \hat{p}_{11})}{\gamma.. + 1} \tag{5.3}$$

$$Var(p_1.) = \frac{\hat{p}_1.(1 - \hat{p}_1.)}{\gamma.. + 1} \tag{5.4}$$

$$Var(p._1) = \frac{\hat{p}._1(1 - \hat{p}._1)}{\gamma.. + 1} \tag{5.5}$$

As for the covariances, the set of fundamental parameters: $\{p_{00}, p_{01}, p_{10}, p_{11}\}$ follows a $Dir(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$ distribution. The general formula for the covariance between p_{00}, p_{01}, p_{10} and p_{11} is [Wil62] (for annotational simplicity, $\hat{p}_{ij}$ here denotes $\gamma_{ij}/\gamma..$ etc.):

$$Cov(p_{ij}, p_{kl}) = \frac{\delta_{ik}\delta_{jl}\hat{p}_{ij} - \hat{p}_{ij}\hat{p}_{kl}}{\gamma.. + 1} \tag{5.6}$$

Based on this, explicit formulae for the covariances between p_{11}, $p_1.$ and $p._1$ can be derived:

$$Cov(p_{11}, p_1.) = Cov(p_{11}, p_{11} + p_{10}) = Var(p_{11}) + Cov(p_{11}, p_{10}) = \tag{5.7}$$

$$= \frac{\hat{p}_{11}(1 - \hat{p}_{11})}{\gamma.. + 1} - \frac{\hat{p}_{11}\hat{p}_{10}}{\gamma.. + 1} = \frac{\hat{p}_{11}(1 - \hat{p}_{11} - \hat{p}_{10})}{\gamma.. + 1} =$$

$$= \frac{\hat{p}_{11}(1 - \hat{p}_1.)}{\gamma.. + 1}$$

and

$$Cov(p_{11}, p_{\cdot 1}) = \frac{\hat{p}_{11}(1 - \hat{p}_{\cdot 1})}{\gamma_{\cdot\cdot} + 1} \tag{5.8}$$

Finally:

$$
\begin{aligned}
Cov(p_{1\cdot}, p_{\cdot 1}) &= Cov(p_{11} + p_{10}, p_{11} + p_{01}) = \\
&= Var(p_{11}) + Cov(p_{11}, p_{10}) + Cov(p_{11}, p_{01}) + Cov(p_{10}, p_{01}) = \\
&= \frac{\hat{p}_{11}(1 - \hat{p}_{11})}{\gamma_{\cdot\cdot} + 1} - \frac{\hat{p}_{11}\hat{p}_{10}}{\gamma_{\cdot\cdot} + 1} - \frac{\hat{p}_{11}\hat{p}_{01}}{\gamma_{\cdot\cdot} + 1} - \frac{\hat{p}_{10}\hat{p}_{01}}{\gamma_{\cdot\cdot} + 1} = \\
&= \frac{\hat{p}_{11}\hat{p}_{00} - \hat{p}_{10}\hat{p}_{01}}{\gamma_{\cdot\cdot} + 1}
\end{aligned}
\tag{5.9}
$$

Note that whereas the sign of $Cov(p_{1\cdot}, p_{\cdot 1})$ varies, neither $Cov(p_{11}, p_{1\cdot})$ nor $Cov(p_{11}, p_{\cdot 1})$ are ever negative. This may explain why the spread of the naïve simulation (based on independent sampling) in Figure 5.2 was remarkably larger than those based on the two Dirichlet based simulations: the positive covariance between the nominator (p_{11}) and the denominator ($p_{1\cdot}$ and $p_{\cdot 1}$ respectively) reduces the spread of the IC distribution (since when p_{11} is unexpectedly high or low, the chances that $p_{1\cdot}$ and $p_{\cdot 1}$ are as well increase). Disregarding this leads to an over-estimation of the IC distribution's variability.

5.4.2 On the equivalence of mD and PG data models under Haldane-like priors

With Haldane-like priors $\{p_{00}, p_{01}, p_{10}, p_{11}\} \sim Dir(0, 0, 0, 0)$ and $\lambda_{00} \sim Ga(0, 0)$ etc., the mD and the PG data models yield equivalent posterior IC distributions.

Consider the PG data model. With counts c_{00} etc. the posterior distributions for the intensities λ_{00} etc. will be: $Ga(c_{00}, 1)$ etc. (see Section 2.1.2). The corresponding probabilities p_{00} etc. are given by the ratios $\frac{\lambda_{00}}{\lambda_{00} + \lambda_{01} + \lambda_{10} + \lambda_{11}}$ etc.

A known relationship between Gamma and Dirichlet distributed variable sets states that for a given set of Gamma distributed variables $\{X_1, \ldots, X_k\}$ where $X_i \sim Ga(x_i, 1)$, the set $\{\frac{X_1}{\sum_i X_i}, \ldots, \frac{X_k}{\sum_i X_i}\}$ is $Dir(x_1, \ldots, x_k)$ distributed. This means that $\{p_{00}, p_{01}, p_{10}, p_{11}\} \sim Dir(c_{00}, c_{01}, c_{10}, c_{11})$, which is exactly equivalent to the posterior distribution in the mD model for these priors.

The similarity between the two methods is further emphasized by the fact that the the marginal counts in the mD model follow binomial distributions, that are asymptotically equivalent to Poisson distributions. In particular, the Poisson distribution with intensity λ is asymptotically equivalent to a binomial distribution of n trials and probability of success p, if n tends to infinity and p to 0 while the product np remains fixed (and equal to λ) [Ric95]. The large sparse databases often considered in IC analysis, feature exactly these characteristics.

Altogether, this indicates that the performance of IC analysis should be rather robust with respect to the choice between the mD and the PG data models.

5.4.3 On the similarity between the regular and the Bayesian bootstrap

This section shows that for multinomially distributed data, again with Haldane-like priors, there is a close relationship between the Bayesian bootstrap method (see Section 2.2.5) and Monte Carlo simulation from the posterior. Because of the equivalence between the mD and the PG data models under these circumstances (see Section 5.4.2), this result also holds for Poisson distributed data.

Given a data set with counts c_{00}, c_{01}, c_{10} and c_{11}, considered to follow a multinomial distribution, assume a joint $Dir(0, 0, 0, 0)$ prior distribution for the corresponding probabilities p_{00}, p_{01}, p_{10} and p_{11}. This gives a corresponding $Dir(c_{00}, c_{01}, c_{10}, c_{11})$ posterior distribution.

The Bayesian bootstrap method assigns $Dir(1, 1, \ldots, 1)$ distributed random weights to the records in the database (see Section 2.2.5). If we, instead of considering the random weights for each individual record in the database, aggreagate the random weights for each record *type*, they will follow a $Dir(c_{00}, c_{01}, c_{10}, c_{11})$ distribution. Consequently, in each Bayesian bootstrap draw, the parameters $\{p_{00}, p_{01}, p_{10}, p_{11}\}$ follow the same distribution as in the regular Monte Carlo simulation from the posterior (based on Haldane-like priors).

The trademark of the Bayesian bootstrap is that it assumes no data model and no prior distribution. However, it seems that Monte Carlo simulation from a posterior distribution based on Haldane-like priors imitates the Bayesian bootstrap—both with the mD and the PG data models.

Chapter 6

Results

The implementation of the new Monte Carlo method has resulted in four main contributions:

- A characterization of the IC distribution and its sensitivity to different parameter values

- An investigation into the validity of different approximations to the IC distribution

- A proposed new approach for derivation of $IC_{\alpha/2}$ estimates

- A comparison of the classical and the Bayesian approach, via the regular and the Bayesian bootstraps

6.1   Characteristics of the IC distribution

In this section, different characteristics of the IC distribution are presented. The study is based on Monte Carlo simulations of the distribution, with 30 000 draws in each simulation. The mD data model described in Section 3.3.3 is assumed to be the correct model of how data is generated, the annotation used in this section was introduced in Section 5.1, and the setup of the experiment and how the parameters of the Dirichlet distribution relate to the counts in the database was explained in Section 5.2.

Also, note that when the impact of a certain parameter is studied, the values of the other (fixed) parameters are somewhat arbitrary. The counts in high dimensional databases (databases with many different variables) are however generally such that c_{11} is much smaller than $c_{1.}$ and $c_{.1}$, and these in turn generally much smaller than $c_{..}$, and this is to some extent reflected in the choice of values used for the fixed γ parameters.

The results presented, apply to both prior and posterior IC distributions, as long as the parameters are modeled by joint Dirichlet distributions. The parameters of the Dirichlet distribution are denoted γ_{00}, γ_{01}, γ_{10} and γ_{11} both
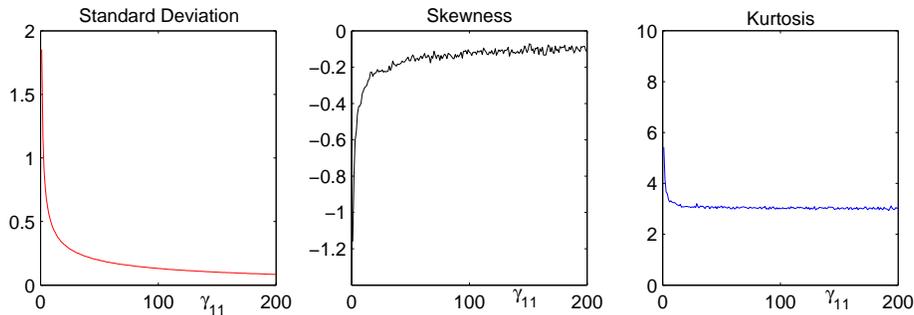
Figure 6.1: The impact of γ_{11} on standard deviation, skewness and kurtosis, for simulated IC distributions with γ_{11} between 1 and 100 and the other parameters constant: $\gamma_{1.} = \gamma_{.1} = 10^3$ and $\gamma_{..} = 10^6$.

for prior and posterior distributions, and $\gamma_{1.}$ refers to the sum of γ_{10} and γ_{11} etc.

6.1.1 The impact of γ_{11}

Figure 6.1 indicates that variation in γ_{11} tends to have a rather significant impact on all three features—standard deviation, skewness and kurtosis. This impact is especially dramatic for low γ_{11} (approximately between 1 and 25, for this particular value of γ_{11}) when the standard deviation drops from 2 to below 0.4 (bits), the skewness increases from -1 to -0.2 and kurtosis drops from 6 to close to 3. The two latter observations indicate that the IC distribution approaches a normal distribution for increased γ_{11}. The skewness however, does not tend to 0 as rapidly as the kurtosis tends to 3, and it is unclear from this diagram whether or not the asymptotic tendency of skewness is to approach 0 or not. However, simulations of IC distributions with $\gamma_{11} = 1000$ give skewness in the order of magnitude 10^{-2} and simulations with $\gamma_{11} = 10\ 000$ give skewness in the order of magnitude 10^{-4}, something that indicates that the skewness does in fact seem to tend to zero. As expected, low values of γ_{11} give the most asymmetric and spread out IC distributions.

To show how the shape of the IC distribution varies for different values of γ_{11}, six specific IC distributions are displayed in Figure 6.2. This may facilitate the interpretation of how standard deviation, skewness and kurtosis vary with γ_{11}. The tendency of both standard deviation and asymmetry to decrease as γ_{11} increases is clear.

6.1.2 The impact of $\gamma_{1.}$ and $\gamma_{.1}$

Since the IC distribution is symmetric with respect to $\gamma_{1.}$ and $\gamma_{.1}$, any results derived for $\gamma_{1.}$ are also valid for $\gamma_{.1}$. Unless otherwise stated, $\gamma_{1.}$ is therefore assumed to be the lower of $\gamma_{1.}$ and $\gamma_{.1}$, to simplify annotation.
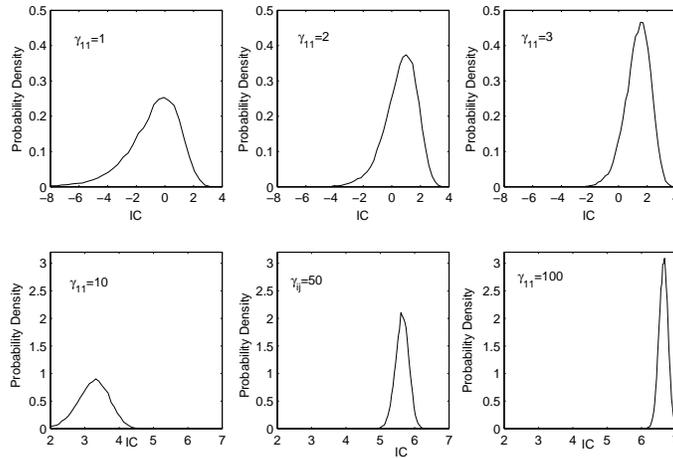
Figure 6.2: The shape of six specific IC distributions with γ_{11} varying between 1 and 100, $\gamma_{1.} = \gamma_{.1} = 10^3$ and $\gamma_{..} = 10^6$.

Figure 6.3 and Figure 6.4 indicate that as long as both $\gamma_{1.}$ and $\gamma_{.1}$ are large compared to γ_{11} (requiring a ratio of at least 10:1 seems to be a good heuristic), the impact of variation in $\gamma_{1.}$ or $\gamma_{.1}$ on the shape of the IC distribution is almost negligible. However, Figure 6.3 also shows that for low values of $\gamma_{1.}$, all three shape features vary significantly. The shape features stabilize when $\gamma_{1.}$ increases above 100: the standard deviation around approximately 0.7 bits, the skewness around -0.6 and the kurtosis around 3.8. The asymptotic values for the shape features with respect to variation in $\gamma_{1.}$ depend heavily on γ_{11}, but there are, for any choice of γ_{11}, asymptotic values of the shape features with respect to variation in $\gamma_{1.}$ or $\gamma_{.1}$.

The decreased standard deviation of the IC distribution when $\gamma_{1.}$ approaches γ_{11} is likely to be attributable to increased covariance between p_{11} and $p_{1.}$, as discussed in Section 5.4.1. Figure 6.5 and Figure 6.6 show that high covariation between p_{11} and $p_{1.}$ coincides with low variance of the IC distribution. The fact that the covariance is proportional to $(1 - \hat{p_{1.}})$, as indicated in Equation 5.7, explains the increased $Cov(p_{11}, p_{1.})$ when $\gamma_{1.}$ decreases.

It is hard to tell whether the impact on skewness and kurtosis for low values of $\gamma_{1.}$, should be attributed to the increased covariation or to the fact that when $\gamma_{1.} \approx \gamma_{11}$, the two parameter values are about equally uncertain.

Figure 6.7 shows three specific IC distributions where the impact of reduced spread when $\gamma_{11} \approx \gamma_{1.}$ is evident. The three distributions have constant values for γ_{11}, $\gamma_{..}$ and for the product $\gamma_{1.} \cdot \gamma_{.1}$ (to fix the horizontal placement), but the values of $\gamma_{1.}$ and $\gamma_{.1}$ vary. The two distributions to the left and in the center, for which $\gamma_{1.} \gg \gamma_{11}$, are very similar, but the right-most distribution has a remarkably lower variance. This is due to the increase in $Cov(p_{11}, p_{1.})$
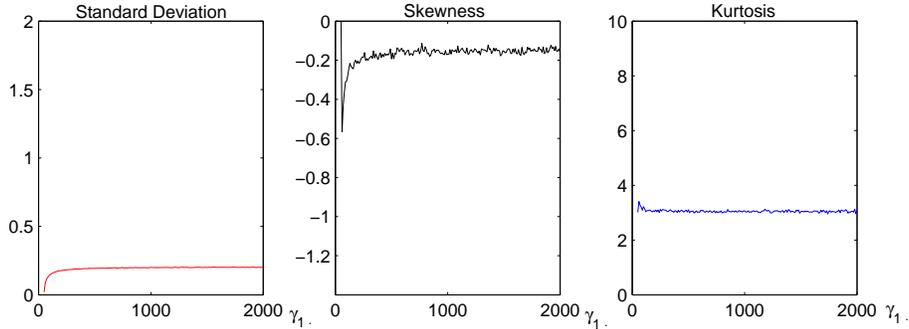
Figure 6.3: Impact on standard deviation, skewness and kurtosis of simulated IC distributions for an increase in the lower of the two marginal parameters (here $\gamma_{1\cdot}$) from 4 to 100. The other parameters are constant: $\gamma_{11} = 4$, $\gamma_{\cdot 1} = 10^3$ and $\gamma_{\cdot\cdot} = 10^6$.

when $\gamma_{1\cdot}$ approaches γ_{11}, that leads to decreased spread of the IC distribution as discussed in Section 5.4.1.

6.1.3 The impact of $\gamma_{\cdot\cdot}$

The parameter $\gamma_{\cdot\cdot}$ appears to have a rather limited impact on the shape of the IC distribution. Figure 6.8 shows that, when the other parameters are fixed, the standard deviation, skewness and kurtosis are robust with respect to variation in $\gamma_{\cdot\cdot\cdot}$. The limited influence of $\gamma_{\cdot\cdot}$ is further illustrated by the fact that the distributions in Figure 6.9 are so similar in shape despite significant variation in $\gamma_{\cdot\cdot\cdot}$.

6.1.4 Combined parameter impact

The most important conclusion of the previous three sections is that when $\gamma_{11} < 0.1 \cdot \min(\gamma_{1\cdot}, \gamma_{\cdot 1})$, the shape of the IC distribution is largely determined by the value of γ_{11} alone. To illustrate this, Figure 6.10 shows six IC distributions with a common γ_{11} value, but varying values for $\gamma_{1\cdot}$, $\gamma_{\cdot 1}$ and $\gamma_{\cdot\cdot\cdot}$. All six distributions appear to be very similar in shape (possibly with the exception of the two distribution in the bottom right corner where $\min(\gamma_{1\cdot}, \gamma_{\cdot 1}) \approx 10c_{11}$). This supports our previous conclusions about the general unsignificant impact of $\gamma_{1\cdot}$, $\gamma_{\cdot 1}$ and $\gamma_{\cdot\cdot}$ on the shape of the IC distribution. Any dissimilarity between these distributions would in contrast have indicated either interaction effects due to the concurrent variation of $\gamma_{1\cdot}$, $\gamma_{\cdot 1}$ and $\gamma_{\cdot\cdot}$ or effects that can not be detected by studying standard deviation, skewness and kurtosis.
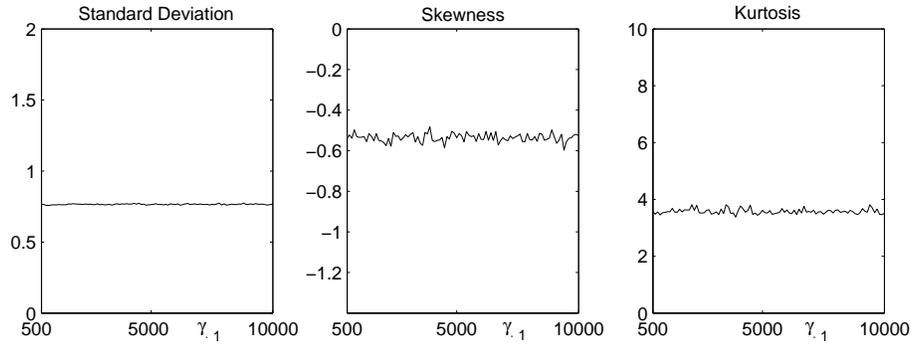
Figure 6.4: Impact on standard deviation, skewness and kurtosis of simulated IC distributions for an increase in the larger of the two marginal parameters (here $\gamma_{\cdot 1}$) from 500 to 10000. The other parameters are constant: $\gamma_{11} = 4$, $\gamma_{1\cdot} = 500$ and $\gamma_{\cdot\cdot} = 10^6$.
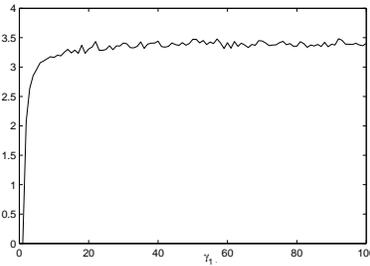
Figure 6.5: The variance of the IC distribution as a function of $\frac{\gamma_{11}}{\gamma_{1\cdot}}$. 100 simulated IC distributions with $\gamma_{11} = 1$, $\gamma_{\cdot 1} = 10^3$, $\gamma_{\cdot\cdot} = 10^6$, and $\gamma_{1\cdot}$ varying between 1 and 100. 30 000 draws per simulation.
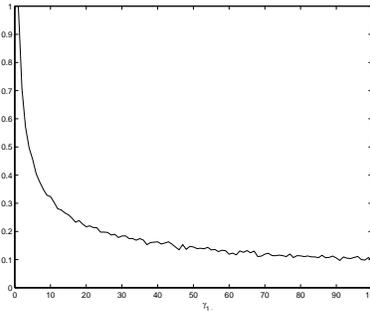
Figure 6.6: The correlation coefficient between p_{11} and $p_{1\cdot}$ as a function of $\frac{\gamma_{11}}{\gamma_{1\cdot}}$. 100 simulated IC distributions with $\gamma_{11} = 1$, $\gamma_{\cdot 1} = 10^3$, $\gamma_{\cdot\cdot} = 10^6$ and $\gamma_{1\cdot}$ varying between 1 and 100. 30 000 draws per simulation.

39
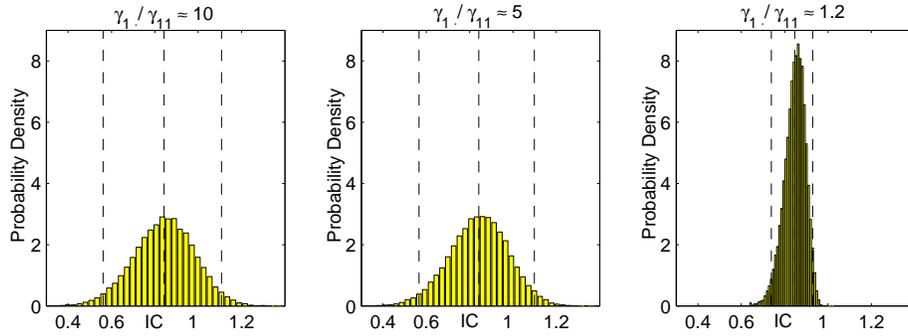
Figure 6.7: Three specific simulated IC distributions with $\gamma_{1\cdot}$ varying from 100 to 1000. Two parameters are fixed: $\gamma_{11} = 90$ and $\gamma_{\cdot\cdot} = 20000$. The parameter and $\gamma_{\cdot 1}$ is adjusted to fix the horizontal displacement.
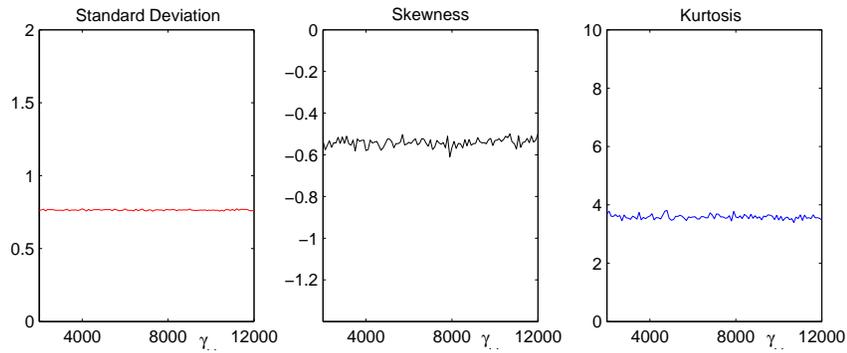
Figure 6.8: The impact on the standard deviation, the skewness and the kurtosis of simulated IC distributions of variation in $\gamma_{\cdot\cdot}$ in steps of 100 from 2000 to 12 000. The other parameters are: $\gamma_{11} = 4$, $\gamma_{1\cdot} = 500$ and $\gamma_{\cdot 1} = 1000$.
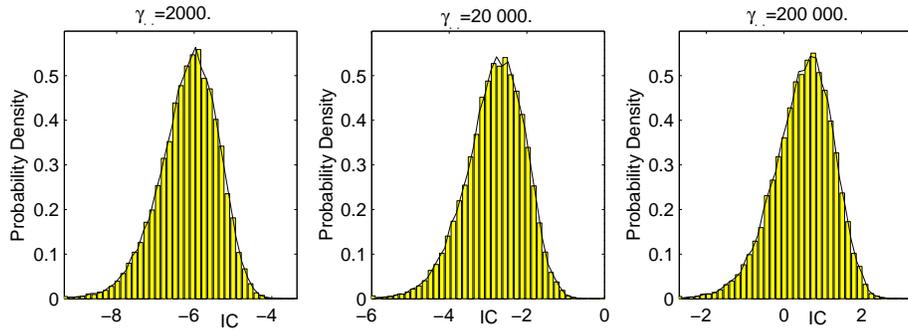
Figure 6.9: Three specific IC distributions with $\gamma_{11} = 3$, $\gamma_{1\cdot} = 500$ and $\gamma_{\cdot 1} = 1000$ in common, and different values of $\gamma_{\cdot\cdot}$.
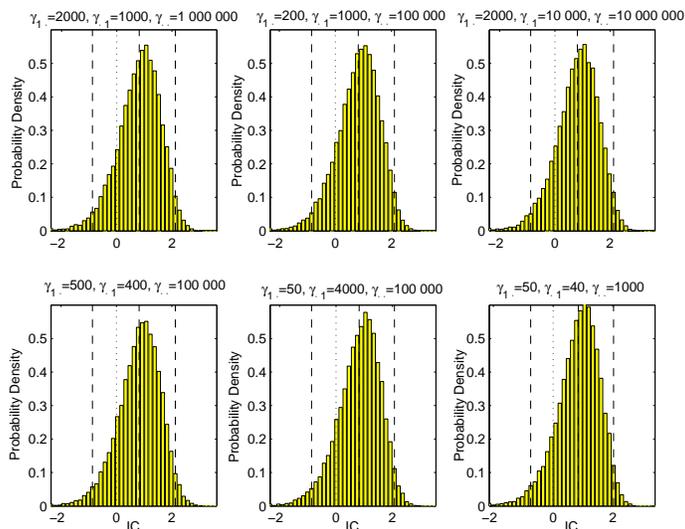
Figure 6.10: Six specific IC distributions with $\gamma_{11} = 4$ and $\frac{\gamma_{..}}{\gamma_{1\cdot}\gamma_{\cdot 1}} = 0.5$ (to fix the horizontal displacement of the distribution). The individual values of $\gamma_{1\cdot}$, $\gamma_{\cdot 1}$ and $\gamma_{..}$ vary.

6.2 Validity of the approximations

This section evaluates the validities of two types of approximate derivation of $IC_{\alpha/2}$: based on the currently implemented normal approximation (see Section 3.5.1), and based on the fixed marginals approximation (see Section 3.5.3). The comparison is made against Monte Carlo based (30 000 draws) $IC_{\alpha/2}$ estimates.

Note that it is the absolute difference between $IC_{\alpha/2}$ estimates that is of interest, since we consider an error of e.g. 0.2 bits to be equally important regardless of how close to 0 the actual estimate is.

Figure 6.11 and Figure 6.15 display the respective accuracies of the $IC_{\alpha/2}$ estimates over wide ranges of parameter values.

6.2.1 Validity of the normal approximation

The results in Section 6.1 indicate that $IC_{\alpha/2}$ estimates based on the current normal approximation will be least appropriate for parameter configurations with low γ_{11} or with $\gamma_{1\cdot}$ or $\gamma_{\cdot 1}$ not significantly larger than γ_{11}. The reason is that for the former type of parameter sets, the true IC distribution is asymmetric and for the latter type of parameter sets, the variance is lower than indicated by the approximate formula for the variance (Equation 3.8).

These predictions are confirmed by Figure 6.11, in which parameter sets that yield inaccurate $IC_{\alpha/2}$ estimates (off by more than 0.2 bits), are indicated (for

41
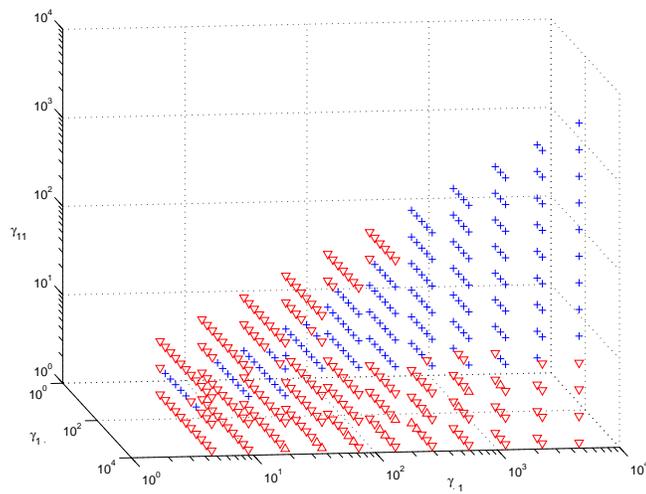
Figure 6.11: This is a three dimensional plot where each point on the graph represents a set of parameter values γ_{11}, $\gamma_{1\cdot}$ and $\gamma_{\cdot1}$. For each parameter set, $IC_{\alpha/2}$ has been estimated both based on Monte Carlo simulation (30 000 draws) and based on the current normal approximation (see Section 3.5.1). Triangles (red) indicate parameter sets for which the two estimates differ by more than 0.2 bits and plus signs (blue) indicate parameter sets for which the two estimates differ by less than 0.2 bits. γ_{11} is varied between 1 and 1024, $\gamma_{1\cdot}$ and $\gamma_{\cdot1}$ are varied between 5 and 5120 and $\gamma_{\cdot\cdot}$ is constant and equal to 1 000 000.
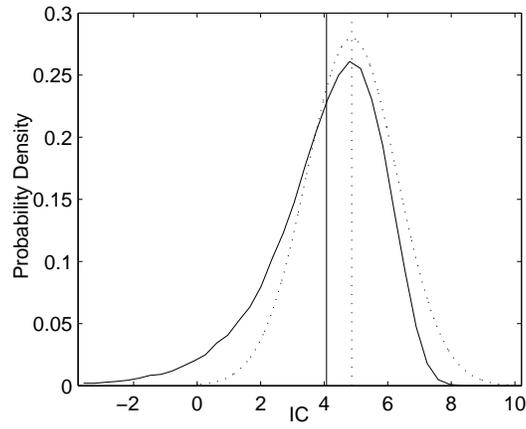
Figure 6.12: Comparison between exact and approximate posterior mean estimates for the IC distribution with parameters $\gamma_{11} = 1$, $\gamma_{1.} = 100$, $\gamma_{.1} = 1000$ and $\gamma_{..} = 2830764$. The dotted distribution is the normal approximation that is currently used.
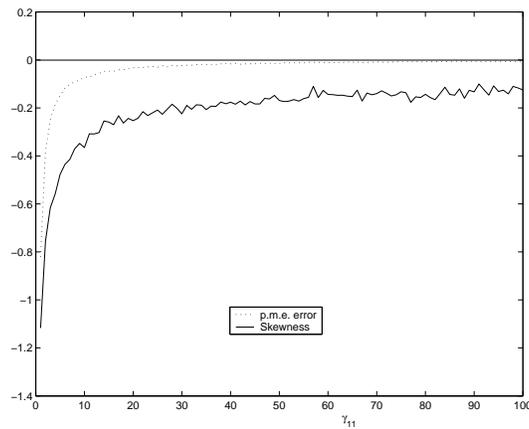
Figure 6.13: The accuracy (compared to simulation) of the current p.m.e. and skewness for different values of γ_{11} (between 1 to 100) and with $\gamma_{1.} = 10^3$, $\gamma_{.1} = 10^3$ and $\gamma_{..} = 10^6$.
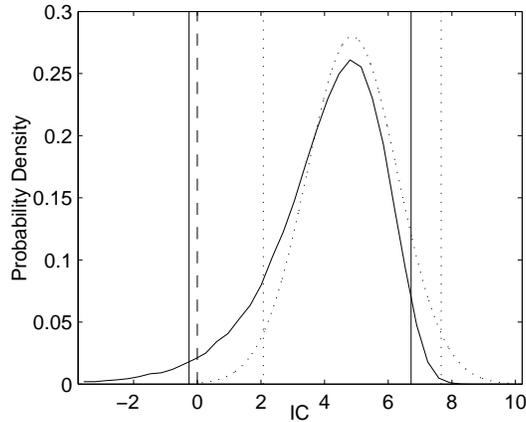
Figure 6.14: Comparison between exact and approximate IC credibility intervals for the asymmetric IC distribution with parameters $\gamma_{11} = 1$, $\gamma_{1.} = 100$, $\gamma_{.1} = 1000$ and $\gamma_{..} = 2830764$. The normal approximation used currently is displayed in dots.

better interpretability, no distinction between positive and negative deviations is made).

The parameter sets yielding inaccurate $IC_{\alpha/2}$ estimates can, as indicated in Figure 6.11, be separated into two intervals (that may not be optimally *compact*):

1. $\gamma_{11} \leq 10$

2. $\gamma_{11} \geq 0.1 \min(\gamma_{1.}, \gamma_{.1})$

A normal approximation, no matter how refined, can only be valid if the distribution at hand is fairly close to normally distributed. One of the standard tests for normality (e.g. the Jarque-Bera test or the Lilliefors test) could be used to determine this. On the other hand, it is clear from inspection of the simulated distributions and plots of skewness and kurtosis, that for low values of γ_{11} the IC distribution is definitely non-normal. In addition, a test for normality does not account for the impact of approximate formulae for the mean and variance.

As pointed out in Section 6.1.1, the IC distribution tends to be heavy-tailed towards the left when it is asymmetric. This will, for parameter sets with low γ_{11}, lead to type I errors: over-estimating the significance of a finding.

That the true covariance is higher than what is indicated by the approximate formula for the variance will on the other hand, for parameter sets with γ_{11} close to $\gamma_{1.}$ or $\gamma_{.1}$, lead to type II errors: under-estimating the significance of an indication.

Figure 6.13 plots the accuracy of the approximate posterior mean estimate (Equation 3.7) and the skewness of the IC distribution against γ_{11}. This fig-
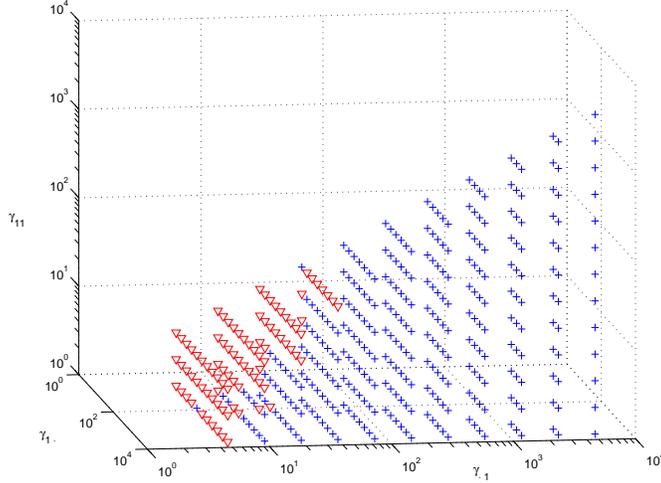
Figure 6.15: This is a three dimensional plot where each point on the graph represents a set of parameter values γ_{11}, $\gamma_{1\cdot}$ and $\gamma_{\cdot 1}$. For each parameter set, $\hat{IC}_{\alpha/2}$ has been estimated both based on Monte Carlo simulation (30 000 draws) and based on the fixed marginals approximation (see Section 3.5.3). Triangles (red) indicate parameter sets for which the two estimates differ by more than 0.2 bits and plus signs (blue) indicate parameter sets for which the two estimates differ by less than 0.2 bits. γ_{11} is varied between 1 and 1024, $\gamma_{1\cdot}$ and $\gamma_{\cdot 1}$ are varied between 5 and 5120 and $\gamma_{\cdot\cdot}$ is constant and equal to 1 000 000.

ure clearly illustrates the tendency of increased skewness and decreased p.m.e. accuracy for low γ_{11}. This effect is important to examine, since an inaccurate p.m.e. approximation displaces the entire estimated distribution, and thereby directly alters $\hat{IC}_{\alpha/2}$. Especially for asymmetric distributions, the inaccuracy of the approximate p.m.e. is large enough to motivate refined calculations. The obvious solution is to replace Equation 3.7 by Equation 3.12.

As discussed in Section 5.4.1 and shown in Figure 6.7, the increased covariance between e.g. p_{11} and $p_{1\cdot}$ as γ_{11} approaches $\gamma_{1\cdot}$ results in a decreased spread of the IC distribution. The current approximate formula for the variance disregards covariance and therefore fails to account for this effect. Consequently, the accuracy of the approximate variance formula decreases as γ_{11} increases relative to $\gamma_{1\cdot}$.

6.2.2 Validity of the fixed marginals approximation

The impact of approximating the IC distribution with a distribution that has fixed marginal probabilities $p_{1\cdot}$ and $p_{\cdot 1}$ is remarkably slight. As indicated in Figure 6.15 the fixed marginals approximation (see Section 3.5.3) of $IC_{\alpha/2}$ gen-

45

erally deviates by less than 0.2 bits for most sets of parameters γ_{11}, $\gamma_{1\cdot}$ and $\gamma_{\cdot 1}$, except for some configurations when $\min(\gamma_{1\cdot}, \gamma_{\cdot 1}) < 100$.

A motivation for the limited impact of assuming $p_{1\cdot}$ and $p_{\cdot 1}$ to be fixed, is that the contribution from uncertainty in each parameter to the variance of the IC distribution is proportional to $\frac{(1-\hat{p})}{\hat{p}}$ (see Equation 3.8). This ratio decreases with increased $\hat{p}$ on the entire interval $[0, 1]$. Therefore, the contribution from the parameter p_{11} to variability in the IC is always larger than the contribution from each of $p_{1\cdot}$ and $p_{\cdot 1}$ (since $p_{11} < \min(p_{1\cdot}, p_{\cdot 1})$). This may explain the unexpectedly high accuracy of the fixed marginals approximation.

6.3   Proposed approach

This section introduces two ways to implement Monte Carlo simulation in IC analysis: brute force simulation and a tabular method. A detailed algorithm for how the different available methods can be combined to yield accurate results in a computationally efficient way is also proposed.

6.3.1   Two Monte Carlo based methods

The straightforward way to implement a Monte Carlo based method for IC analysis is to run Monte Carlo simulations for each investigated parameter set. However, in a database where there are hundreds of thousands, or more, IC distributions to investigate, it is usually computationally intractable to run Monte Carlo simulation for every single distribution.

Alternatively, a tabular method can be used. This is an approximate method based on Monte Carlo simulation that resembles how traditional statistical tables are used. A naïve such approach would be to try to tabulate $IC_{\alpha/2}$ values for all possible parameter configurations, but this would for most databases be futile: in dynamical databases $c_{\cdot\cdot}$ changes constantly and so would consequently $\gamma_{\cdots}$. The strategy proposed here instead utilizes the fact that the shape of the IC distribution is generally invariant to variation in all parameters but γ_{11}. This means that for a given γ_{11}, the distance between IC_{pme} and $IC_{\alpha/2}$ will be constant. The idea is to tabulate IC_{Δ} ($= IC_{pme} - IC_{\alpha/2}$) values based on precise Monte Carlo simulations for common values of γ_{11}. Then, for any given parameter configuration, $IC_{\alpha/2}$ can be approximated by $\hat{IC}_{pme}(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}) - IC_{\Delta}(\gamma_{11})$. This approach is particularly useful if only a limited set of values for γ_{11} is observed in the database, and accurate enough if $\min(\gamma_{1\cdot}, \gamma_{\cdot 1}) > 10\gamma_{11}$

6.3.2   Summary of available methods

In summary, there are now four methods for $\hat{IC}_{\alpha/2}$ derivation: the normal approximation (see Section 3.5.1), the fixed marginals approximation (see Section 3.5.3), brute force simulation and the tabular method. These approaches vary in running time and the domain of parameter values for which they give
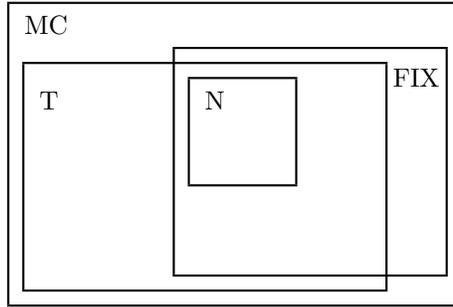
Figure 6.16: A Venn diagram that indicates how the applicability ranges of the four different methods relate. The abbreviations are MC: brute force simulation, T: the tabular method, FIX: the fixed marginals approximation and N: the normal approximation.

accurate results (in this discussion, a method that gives $IC_{\alpha/2}$ estimate with an accuracy of 0.2 bits for a certain parameter set is considered to be applicable to that parameter set):

The current normal approximation Applicable when $\gamma_{11} > 10$ and $\min(\gamma_{1\cdot}, \gamma_{\cdot 1}) > 10\gamma_{11}$. Evaluation of two simple formulae—very fast.

The fixed marginals approximation Applicable when $\min(\gamma_{1\cdot}, \gamma_{\cdot 1}) > 100$. Requires a Newton-Raphson iteration.

The tabular method Applicable when $\min(\gamma_{1\cdot}, \gamma_{\cdot 1}) > 10\gamma_{11}$ and the γ_{11} value of interest is tabulated. Requires the calculation of one series expansion and the retrieval of a tabulated value at execution, but there is a significant setup time that increases linearly with the number of precalculated IC_Δ values.

Brute force simulation Applicable over the entire domain, but computationally intensive and the complexity grows with required accuracy.

6.3.3   Proposed algorithm

If time is limited, complex Monte Carlo simulations should be avoided to as large an extent as possible. Figure 6.16 displays a Venn diagram that illustrates how the validity domains of the four different methods relate. Note that any time the normal approximation is applicable, so is the fixed marginals approximation, and that the opposite is not true.

The recommended method is therefore to (the first step is carried out once and for all initially, but the three following steps are carried out for every parameter set investigated):

1. Run Monte Carlo simulations (30 000 draws) and store IC_Δ values for the most common γ_{11}

2. If $\min(\gamma_{1.}, \gamma_{.1}) > 10\gamma_{11}$ and γ_{11} is stored, then use the tabular method

3. Else if $\min(\gamma_{1.}, \gamma_{.1}) > 100$, then use the fixed marginals method

4. Else run a specific Monte Carlo simulation (30 000 draws)

6.3.4 Tractability of the proposed algorithm

To give an idea of the proportion of combinations that could be handled with each approach in the proposed method (see Section 6.3) when applied to the WHO database, we consider the database after the second quarterly update, 2002 (assume that there are stored IC_Δ values for γ_{11} between 1 and 1000). This table displays two numbers: the number of drug/ADR combinations that each approach would be applicable to (expected accuracy better than 0.2 bits), and the number of combinations that would actually be handled by each type analysis if the proposed algorithm was applied to the data set:

Approach	Applicable to:	Applied to in proposed algorithm:
Brute force simulation	571 685 (100%)	42 639 (7.5%)
Tabular method	526 170 (92%)	526 170 (92%)
Fixed marginals approximation	439 310 (77%)	2 876 (0.5%)
Normal approximation	77 403 (14%)	0 (0%)

It is worth noting that the reason for the limited applicability of the normal approximation is that over 85% of the observed drug/ADR combinations in the WHO database have counts smaller than 10. Also note that the fixed marginals approach is accurate for very few combinations that cannot be handled by the tabular approach.

To run the 43 000 Monte Carlo simulations with 30 000 draws each, is estimated to take less than 4 hours on a computer with a 1 GHz processor and the algorithms implemented in the C programming language (Roland Orre, personal communication).

6.4 A classical approach: bootstrapping the IC distribution

One way to determine the impact of the Bayesian approach on IC analysis, is to compare Bayesian and regular bootstrap distributions for the IC. This was carried out as part of this thesis project for an artificial data set with aggregated counts: $c_{11} = 3$, $c_{1.} = 20$, $c_{.1} = 100$, and $c_{..} = 1000$.
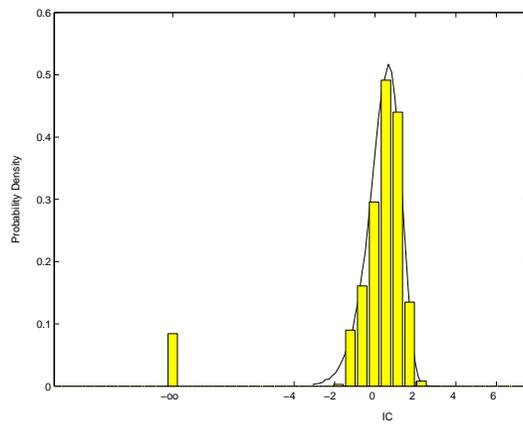
Figure 6.17: Comparison of the results of a Bayesian and a regular bootstrap distribution estimate. 100 000 draws in each simulation from a batch with $c_{1\cdot} = 20$, $c_{\cdot 1} = 100$, $c_{11} = 3$ and $c_{\cdot\cdot} = 1000$. The histogram indicates the result of the regular bootstrap, and the line indicates the result of the Bayesian bootstrap.
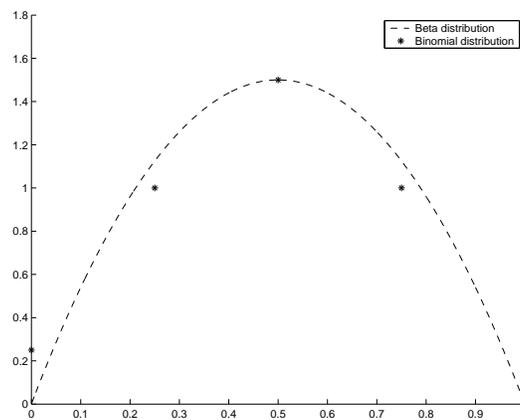
Figure 6.18: Comparison of Be(2,2) to a scaled bin(4,0.5).
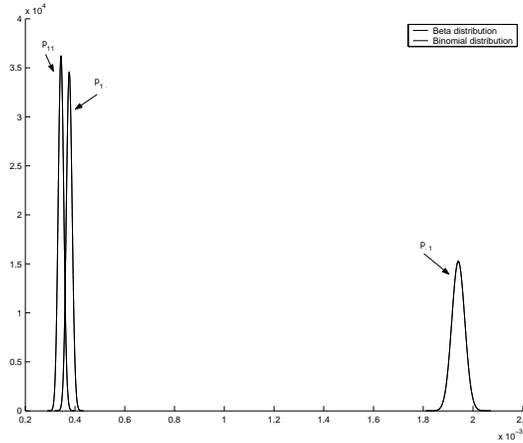
Figure 6.19: Comparison between the Beta distributions and their corresponding (scaled) binomials for an authentic combination in the database. The counts are rather large: $c_{1.} = 5495$, $c_{.1} = 1067$, $c_{11} = 973$ and $c_{..} = 2830764$. There *are* slight differences between the three distribution pairs although they are almost impossible to make out in this resolution.

First note, that for the problem at hand (probability estimation from dichotomous data) the parametric and the non-parametric versions of the regular bootstrap are equivalent—they both correspond to sampling from a binomial distribution [Hjo94]. Therefore, the comparison to the Bayesian bootstrap automatically applies to both.

Figure 6.17 displays the result of the regular and the Bayesian bootstrap distributions. They are rather similar, apart from the $-\infty$ peak of the regular bootstrap distribution. This is due to the rare but possible $\hat{p}^*_{11} = 0$ simulated parameter estimate that results in $IC^* = \log 0$ in the regular bootstrap. $\hat{p}_{11} = 0$ estimates occur when no record with both states is selected in the resampling procedure (see Section 2.2.2).

Another important difference between the regular and the Bayesian bootstrap methods is that the Bayesian bootstrap may produce any IC value, but that the regular bootstrapped IC distribution is discrete (the simulated $\hat{p}_{11}$, $\hat{p}_{1.}$ and $\hat{p}_{.1}$ are always multiples of $\frac{1}{c_{..}}$).

Now, consider the regular bootstrap distribution of one of the constituting parameter estimates, say $\hat{p}_{11}$, and the Bayesian bootstrap distribution of the corresponding parameter p_{11}. The parameter estimate $\hat{p}_{11}$ is essentially binomially distributed ($\hat{p}_{11} = \frac{c_{11}}{c_{..}}$, $c_{11} \sim bin(c_{..}, p_{11})$) and the parameter p_{11} is Beta distributed ($p_{11} \sim Be(c_{11}, c_{..} - c_{11})$. Consequently any difference between the classical bootstrap and the Bayesian bootstrap is to a large extent due to differences between the $Be(\alpha, \beta)$-type distributions of the constituting parameters and the scaled $bin(\alpha + \beta, \frac{\alpha}{\alpha+\beta})$-type distribution of the parameter estimates.

50

Comparisons between Beta distributions and their corresponding scaled binomial distributions for two sample sets of parameters are presented in Figure 6.18 and Figure 6.19. Figure 6.18 illustrates that for low enough $\min(\alpha, \beta)$, the two distributions are clearly dissimilar. Figure 6.19, on the other hand shows that when $\min(\alpha, \beta)$ is large enough for all three parameters, the two distributions tend to coincide.

In summary, we see that the two approaches differ for the low counts that are typical for the WHO database, and that the main advantages of the Bayesian bootstrap compared to its regular correlate, is that the simulated distribution is continuous and that $IC^* \neq -\infty$.

Chapter 7

Discussion

7.1 General issues

We first discuss some general issues in statistical inference that are particularly important to any attempt at knowledge discovery.

7.1.1 Multiple Comparisons Issues

When several tests are performed in parallel, the significance of each finding depends on the total number of tests carried out. The reason for this is obvious: if several equivalent tests are carried out, the probability that at least one of these tests will deviate significantly from the expected, is larger than the same probability for each individual test. The impact of this effect may be striking, even for a moderate number of tests. Consider for example the probability that at least one out of 100 parallel tests deviates significantly at the 1% level! This probability is as high as $1 - 0.99^{100} = 63\%$

This severely impairs any statement about the significance of findings in explorative statistical analysis (e.g. data mining applications such as whole genome scans). There are statistical methods designed to handle this, for example the Bonferroni method and Tukey's methods for multiple comparisons [Ric95]. An alternative approach is to reserve a subset of the data, and use it for follow up studies aimed to confirm any indications from the initial explorative study.

However, when IC analysis is used merely as a means to rank possible patterns in a database, the multiple comparisons aspect is not very important. In this case, the purpose is not to prove that an indication is significant, but to highlight the dependencies that are most interesting for further research. Since all dependencies are equally affected by the multiple comparisons aspect, the ranking is not biased.

7.1.2 Confounding variables

A confounder is a non-controlled variable that accounts for some of the variation in one or several of the controlled variables. Some confounders are automatically handled properly in IC analysis. One such example from the application to the WHO database is the variable prescription rates of different drug substances. If a certain drug substance has an overall high but even report rate for all different ADR's, this may indicate that this drug substance is widely prescribed and consequently more likely to be erroneously suspected of causing ADR's. As desired, this does not affect the Information Components related to this drug substance, since both p_{11} to $p_{1.}$ are affected by the higher prevalence.

Another type of confounder is the "common cause" variable. It is a non-controlled variable that is the common cause of two observed variables. This phenomenon may lead to significant over-estimation of the strength of a certain dependency. An example from drug monitoring is the dependency in the database between the polio vaccine and the sudden infant death syndrome (s.i.d.s.). This is due simply to the fact that both these factors are correlated with young age: polio vaccine is generally given to infants, and all victims of s.i.d.s. are per definition infants. If the database is stratified with respect to different age groups, the dependency is insignificant in each stratum [DuM99].

The phenomenon of an indication being weaker in both sub populations than in the pooled population, is a variation on Simpson's paradox. Simpson's paradox refers to the contra-intuitive (but nevertheless fully logical) observation, that aggregating two separate populations may reverse a common tendency of the two individual populations. For a general description of Simpson's paradox, see a standard statistical inference text book such as [Ric95].

7.2 Comments to the results

There is an infinite number of possible parameter configurations, and the conclusions in this report with respect to the shape of the IC distribution and the validity of the approximate formulae are based on the assumption that the IC distribution is well-behaved enough that observed trends and tendencies extrapolate/interpolate properly to non-investigated parameter combinations. This corresponds to a general assumption that the impact of variation in one parameter is not affected by the values of the other parameters. One exception to this rule is known: the impact of $\gamma_{1.}$ and $\gamma_{.1}$ depend on the value of γ_{11}. Any other such pairwise interactions are however likely to have been been identified in the systematic testing setup that is used. On the other hand, this setup may well miss any interactions between three or four of the parameters, if they exist. Under all circumstances, it seems to be a fair assumption that in the domain where variation of $\gamma_{1.}$, $\gamma_{.1}$ or $c_{..}$ alone, has no significant impact on the shape of the IC distribution, simultaneous variation should not either.

It may be argued that the results of the systematic evaluation on which this report is based is optimized for parameter values typical for high dimensional

databases (in particular, big differences in magnitude between γ_{11} and $\gamma_{..}$), and that the results should be further evaluated before applied to databases where this assumption does not hold.

The use of standard deviation, skewness and kurtosis to quantitatively compare the shapes of different distributions is not perfectly distinctive: two distributions may have the exact same values for these features even though the actual distributions are different. On the other hand, simulations from two very similar distributions are unlikely to yield significantly differing feature estimates. Consequently, variation in any of these feature estimates generally indicates variation in the shape of the distribution.

Section 6.2 shows that the currently used approximate formula for the p.m.e. is inaccurate for asymmetric IC distributions. This affects all statements about the accuracy of the current normal approximation in this report. Exact p.m.e. calculation would improve the general accuracy of the normal approximation. Certainly, however, the $IC_{\alpha/2}$ estimate would still be inaccurate for some configurations, due to the asymmetry of the true IC distribution and the disregarded covariance terms in the approximate variance formula.

With a required accuracy of 0.2 bits, the normal approximation turns out to be applicable to a very small proportion of the combinations in the WHO database (14%). The explanation for this is that although the normal approximation is accurate over a large domain of possible parameter sets, it is generally inaccurate for small values of γ_{11}, when the true IC distribution is asymmetric. Unfortunately, low γ_{11} values are fairly common in this and many other of the interesting data sets.

Section 6.1 showed that over a surprisingly large domain of parameter sets, the parameter γ_{11} alone, determines the shape of the IC distribution. This may motivate why the fixed marginals approximation proposed in Section 3.5.3 is accurate over such a large number of parameter sets. The main advantage of the fixed marginals method is that it does not require symmetric distributions, and therefore generally handles low γ_{11} values better.

The proposed tabular method introduced in Section 6.3 has the advantage of executing fast and giving accurate results over a large domain of parameter values. It does however require a certain setup time, during which precise simulations are run for common values of γ_{11}. The tabular method is ideal for databases with a small set of often re-occurring γ_{11} values, like the WHO database where 85% of the non-zero γ_{11} are in the range 1 to 10.

7.3   Future research and development

There are several areas of interest for further research and development of the method:

- The possible implementation of an empirical Bayes method for derivation of reliable informative prior distributions

- The general impact of different informative prior distributions

- Higher order Information Components such as that of order three:

$$IC = \log_2 \frac{p_{111}}{p_{1..}p_{.1.}p_{..1}}$$

- Automatic identification of confounding variables, and the use of stratified or pooled IC estimates

- Extensions of the results presented in this report to the propagation of probability distributions in a Bayesian neural network framework

- Exact formulae for the covariance between e.g. $\log(p_{11})$ and $\log(p_{1.})$

- Thorough investigation of the accuracy of the tabular method for different parameter sets

- Further investigation of the accuracy of the fixed marginals approximation

- A closed form expression for the accuracy of the Monte Carlo-based $\hat{IC}_{\alpha/2}$, as a function of the number of draws

7.4 Conclusions

Information on the true shape of the IC distribution is crucial to proper IC analysis. This report shows that Monte Carlo simulation is a useful approach for such investigations. The results of Monte Carlo simulations show that (as expected, see [BLE$^+$98]) the IC distribution is asymmetric with a large spread, for low parameter values γ_{11}. As γ_{11} increases, the standard deviation of the IC distribution decreases, and the skewness and kurtosis values tend asymptotically to 0 and 3 respectively—the values typical for a normal distribution. Over a surprisingly large domain of parameter sets, the parameter γ_{11} alone, determines the shape of the IC distribution. In particular, when none of the other parameters are smaller than 100, or less than 10 times as large as γ_{11}, neither $\gamma_{1.}$, $\gamma_{.1}$ nor $\gamma_{..}$ affect the shape of the IC distribution (only the horizontal placement).

Furthermore, this report shows that the $IC_{\alpha/2}$ estimates based on the current normal approximation are inaccurate for certain sets of parameters γ_{11}, $\gamma_{1.}$, $\gamma_{.1}$ and $\gamma_{..}$. This applies particularly to the low values of γ_{11}, that are typical in the high dimensional databases that are often subject to IC analysis. This motivates the development of refined methods for $\hat{IC}_{\alpha/2}$ derivation.

Monte Carlo simulation provides accurate $IC_{\alpha/2}$ estimates for all possible parameter sets and this report suggests that Monte Carlo methods be incorporated in routine IC analysis. Brute force simulation is however computationally complex, and therefore a more tractable tabular method is proposed, where the results of pre-run Monte Carlo simulations for common parameter sets are reused.

Bibliography

[AMS⁺96] Rakesh Agrawal, Hiekki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inkeri Verkamo, *Fast discovery of association rules*, Advances in knowledge discovery and data mining, American Association for Artificial Intelligence, 1996, pp. 307–328.

[AY98] Charu C. Aggarwal and Philip S. Yu, *A new framework for item set generation*, PODS 1998, 1998, pp. 18–24.

[BLE⁺98] A. Bate, M. Lindquist, I. R. Edwards, S. Olsson, R. Orre, A. Lansner, and R. M. De Freitas, *A bayesian neural network method for adverse drug reaction signal generation*, European Journal of Clinical Pharmacology (1998), no. 54, 315–321.

[Che78] R. C. H. Cheng, *Generating beta variates with nonintegral shape parameters*, Communications of the ACM **21** (1978), no. 4, 317–322.

[CL01] M. Clyde and H.K. Lee, *Bagging and the bayesian bootstrap*, Artificial Intelligence and Statistics 2001 (T. Richardson and T. Jaakkola, eds.), 2001, pp. 169–174.

[CT91] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, Wiley Series in Telecommunications, John Wiley & Sons, New York, NY, USA, 1991.

[DP01] William DuMouchel and Daryl Pregibon, *Empirical bayes screening for multi-item associations*, Knowledge Discovery and Data Mining, 2001, pp. 67–76.

[DuM99] William DuMouchel, *Bayesian data mining in large frequency tables, with an application to the fda spontaneous reporting systems*, American Statistician **53** (1999), 177–202.

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, *The kdd process for extracting useful knowledge from volumes of data*, Communications of the ACM **39** (1996), no. 11, 27–34.

[Han98] David J. Hand, *Data mining: Statistics and more?*, The American Statistician **52** (1998), 112–118.

[Hea97]   Michael T. Heath, *Scientific computing: An introductory survey*, McGraw-Hill, 1997.

[Häg02]   Olle Häggström, *Finite markov chains and algorithmic applications*, Cambridge University Press, 2002.

[HJN89]   Sture Holm, Peter Jagers, and Olle Nerman, *Statistisk slutledning*, Chalmers tekniska högskola och Göteborgs universitet, 1989.

[Hjo94]   Urban Hjort, *Computer intensive statistical methods*, 1 ed., Chapman & Hall, 1994.

[HTF01]   T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*, Springer, 2001.

[Hut01]   Marcus Hutter, *Distribution of mutual information*, Tech. Report IDSIA-13-01, IDSIA, Corso Elvezia 36, 6900 Lugano, Switzerland, June 2001.

[KO98]   Timo Koski and Roland Orre, *Statistics of the information component in bayesian neural networks*, Tech. Report TRITA-NA-9806, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden, 1998.

[LE93]   M. Lindquist and I.R. Edwards, *Adverse drug reaction reporting in europe: some problems of comparison*, International Journal of Risk & Safety in Medicine **4** (1993), 35–46.

[Lee97]   Peter M. Lee, *Bayesian statistics: An introduction*, 2 ed., Arnold, 1997.

[LT79]   RD Levine and M Tribus (eds.), *The maximum entropy principle*, MIT Press, Cambridge, Massachusetts, 1979.

[Mar70]   J.S. Maritz, *Empirical bayes methods*, Methuen and Co, 1970.

[Mit97]   Tom M. Mitchell, *Machine learning*, 1 ed., McGraw-Hill, 1997.

[OE00]   S. Olsson and I. R. Edwards, *The who international drug monitoring programme*, Side Effects of Drugs Annual (2000), 524–529.

[OLBL00]   Roland Orre, Anders Lansner, Andrew Bate, and Marie Lindquist, *Bayesian neural networks with confidence estimations applied to data mining*, Computational Statistics & Data Analysis **34** (2000), 473–493.

[Raw88]   M.D. Rawlins, *Spontaneous reporting of adverse drug reactions. i: the data*, British Journal of Clinical Pharmacology **1** (1988), no. 26, 1–5.

[Ric95]   John A. Rice, *Mathematical statistics and data analysis*, 2 ed., Duxbury Press, 1995.

[Rub81]   Donald B. Rubin, *The bayesian bootstrap*, Annals of Statistics **9** (1981), no. 1, 130–134.

[Sch81]   Bruce Schmeiser, *Random variate generation*, Proceedings of the 13th conference on Winter simulation, 1981, pp. 227–242.

[Sha48]   C.E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal **27** (1948), 379–423 and 623–656.

[Vea02]   Richard D. De Veaux, *Data mining: a view from down in the pit*, STATS (2002), no. 34, 3–9.

[Wil62]   S.S. Wilks, *Mathematical statistics*, John Wiley & Sons, 1962.

Appendix A

Review of probability distributions

This appendix reviews some of the probability distributions relevant to this thesis.

A.1 The Binomial distribution

The number of successes x out of n independent identically distributed (i.i.d.) trials follows the binomial distribution. The probability of success is denoted p.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \tag{A.1}$$

A.2 The Poisson distribution

The Poisson distribution can be seen as the limit of a binomial distribution as p tends to zero and n to infinity, while the product $np = \lambda$ remains constant. The intensity λ is the unique parameter of the Poisson distribution, whose frequency function is:

$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{A.2}$$

The number of radioactive decays over a fixed period of time is routinely modelled as a Poisson distribution. Another possible application of the Poisson distribution is to model the white blood cell count in a fixed volume of blood.

The Poisson distribution is generally appropriate when the variable to be modelled is a discrete number of events on a continuous 'interval' (typically in space or time).

A.3　The Multinomial distribution

The multinomial distribution is a generalization of the binomial distribution where, in each trial, several different outcomes are possible. It is a discrete probability distribution, that assigns probabilities to different sets of counts $n_1, n_2, \ldots, n_k$, from n i.i.d. trials, where the k different outcome classes occur with respective probabilities $p_1, p_2, \ldots, p_k$.

$$P(n_1, n_2, \ldots, n_k) = \binom{n}{n_1 n_2 \ldots n_k} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k} \qquad (A.3)$$

A.4　The Beta distribution

The Beta distribution's probability density function is:

$$f(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \qquad (A.4)$$

The most important use of the Beta distribution is in Bayesian statistics as the conjugate prior (see Section 2.1.2) to the binomial distribution. It is non-zero only on the $[0, 1]$ interval.

A.5　The Gamma distribution

The Gamma distribution family includes both the exponential and the chi-squared distributions. Its general form is:

$$f(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \qquad (A.5)$$

Gamma distributions are conjugate priors (see Section 2.1.2) to the Poisson distribution.

A.6　The Dirichlet distribution

The Dirichlet distribution's probability density function is:

$$f(p_1, p_2, \ldots, p_k) = \frac{\Gamma(n_1 + n_2 + \cdots + n_k)}{\Gamma(n_1)\Gamma(n_2)\ldots\Gamma(n_i)} p_1^{n_1-1} p_1^{n_2-1} \ldots p_k^{n_k-1} \qquad (A.6)$$

The Dirichlet distribution is the conjugate prior (see Section 2.1.2) to the multinomial distribution.